CALCULUS ILLUSTRATED

VOLUME 3: INTEGRAL CALCULUS

PETER SAVELIEV

To the student

Mathematics is a science. Just as the rest of the scientists, mathematicians are trying to understand how the Universe operates and discover its laws. When successful, they write these laws as short statements called "theorems". In order to present these laws conclusively and precisely, a dictionary of the new concepts is also developed; its entries are called "definitions". These two make up the most important part of any mathematics book.

This is how definitions, theorems, and some other items are used as building blocks of the scientific theory we present in this text.

Every new concept is introduced with utmost specificity.

Definition 0.0.1: square root

Suppose a is a positive number. Then the *square root* of a is a positive number x, such that $x^2 = a$.

The term being introduced is given in *italics*. The definitions are then constantly referred to throughout the text.

New symbolism may also be introduced.

Square root

$$\sqrt{a}$$

Consequently, the notation is freely used throughout the text.

We may consider a specific instance of a new concept either before or after it is explicitly defined.

Example 0.0.2: length of diagonal

What is the length of the diagonal of a 1×1 square? The square is made of two right triangles and the diagonal is their shared hypotenuse. Let's call it a. Then, by the *Pythagorean Theorem*, the square of a is $1^2 + 1^2 = 2$. Consequently, we have:

$$a^2 = 2\,.$$

We immediately see the need for the square root! The length is, therefore, $a = \sqrt{2}$.

You can skip some of the examples without violating the flow of ideas, at your own risk.

All new material is followed by a few little tasks, or questions, like this.

Exercise 0.0.3

Find the height of an equilateral triangle the length of the side of which is 1.

The exercises are to be attempted (or at least considered) immediately.

Most of the in-text exercises are not elaborate. They aren't, however, entirely routine as they require understanding of, at least, the concepts that have just been introduced. Additional exercise *sets* are placed in the appendix as well as at the book's website: calculus123.com. Do not start your study with the exercises! Keep in mind that the exercises are meant to test – indirectly and imperfectly – how well the *concepts* have been learned.

There are sometimes words of caution about common mistakes made by the students.

Warning!

In spite of the fact that $(-1)^2 = 1$, there is only one square root of 1, $\sqrt{1} = 1$.

The most important facts about the new concepts are put forward in the following manner.

Theorem 0.0.4: Product of Roots

For any two positive numbers a and b, we have the following identity:

$$\boxed{\sqrt{a} \cdot \sqrt{b} = \sqrt{a \cdot b}}$$

The theorems are constantly referred to throughout the text.

As you can see, theorems may contain formulas; a theorem supplies limitations on the applicability of the formula it contains. Furthermore, every formula is a part of a theorem, and using the former without knowing the latter is perilous.

There is no need to memorize definitions or theorems (and formulas), initially. With enough time spent with the material, the main ones will eventually become familiar as they continue to reappear in the text. Watch for words "important", "crucial", etc. Those new concepts that do not reappear in this text are likely to be seen in the next mathematics book that you read. You need to, however, be aware of all of the definitions and theorems and be able to find the right one when necessary.

Often, but not always, a theorem is followed by a thorough argument as a justification.

Proof.

Suppose $A = \sqrt{a}$ and $B = \sqrt{b}$. Then, according to the definition, we have the following:

$$a = A^2 \quad \text{and} \quad b = B^2 \ .$$

Therefore, we have:

$$a \cdot b = A^2 \cdot B^2 = A \cdot A \cdot B \cdot B = (A \cdot B) \cdot (A \cdot B) = (AB)^2 \ .$$

Hence, $\sqrt{ab} = A \cdot B$, again according to the definition.

Some proofs can be skipped at first reading.

Its highly detailed exposition makes the book a good choice for *self-study*. If this is your case, these are my suggestions.

While reading the book, try to make sure that you understand new concepts and ideas. Keep in mind, however, that some are more important that others; they are marked accordingly. Come back (or jump forward) as needed. Contemplate. Find other sources if necessary. You should not turn to the exercise sets until you have become comfortable with the material.

What to do about exercises when solutions aren't provided? First, use the examples. Many of them contain a problem – with a solution. Try to solve the problem – before or after reading the solution. You can also find exercises online or make up your own problems and solve them!

I strongly suggest that your solution should be thoroughly *written*. You should write in complete sentences, including all the algebra. For example, you should appreciate the difference between these two:

$$\text{Wrong:} \quad \boxed{\begin{array}{l} 1+1 \\ 2 \end{array}} \qquad \text{Right:} \quad \boxed{\begin{array}{l} 1+1 \\ =2 \end{array}}$$

The latter reads "one added to one is two", while the former cannot be read. You should also justify all your steps and conclusions, including all the algebra. For example, you should appreciate the difference between these two:

Wrong: $\boxed{\begin{array}{l} 2x = 4 \\ x = 2 \end{array}}$ Right: $\boxed{\begin{array}{l} 2x = 4\,; \text{ therefore,} \\ x = 2\,. \end{array}}$

The standards of thoroughness are provided by the examples in the book.

Next, your solution should be thoroughly *read*. This is the time for self-criticism: Look for errors and weak spots. It should be re-read and then rewritten. Once you are convinced that the solution is correct and the presentation is solid, you may show it to a knowledgeable person for a once-over.

Next, you may turn to modeling projects. Spreadsheets (Microsoft Excel or similar) are chosen to be used for graphing and modeling. One can achieve as good results with packages specifically designed for these purposes, but spreadsheets provide a tool with a wider scope of applications. Programming is another option.

Good luck!

June 23, 2020

To the teacher

The bulk of the material in the book comes from my lecture notes.

There is little emphasis on closed-form computations and algebraic manipulations. I do think that a person who has never integrated by hand (or differentiated, or applied the quadratic formula, etc.) cannot possibly understand integration (or differentiation, or quadratic functions, etc.). However, a large proportion of time and effort can and should be directed toward:

- understanding of the concepts and
- modeling in realistic settings.

The challenge of this approach is that it requires more abstraction rather than less.

Visualization is the main tool used to deal with this challenge. Illustrations are provided for every concept, big or small. The pictures that come out are sometimes very precise but sometimes serve as mere metaphors for the concepts they illustrate. The hope is that they will serve as visual "anchors" in addition to the words and formulas.

It is unlikely that a person who has never plotted the graph of a function by hand can understand graphs or functions. However, what if we want to plot more than just a few points in order to visualize curves, surfaces, vector fields, etc.? Spreadsheets were chosen over graphic calculators for visualization purposes because they represent the shortest step away from pen and paper. Indeed, the data is plotted in the simplest manner possible: one cell - one number - one point on the graph. For more advanced tasks such as modeling, spreadsheets were chosen over other software and programming options for their wide availability and, above all, their simplicity. Nine out of ten, the spreadsheet shown was initially created from scratch in front of the students who were later able to follow my footsteps and create their own.

About the tests. The book isn't designed to prepare the student for some preexisting exam; on the contrary, assignments should be based on what has been learned. The students' understanding of the concepts needs to be tested but, most of the time, this can be done only indirectly. Therefore, a certain share of routine, mechanical problems is inevitable. Nonetheless, no topic deserves more attention just because it's likely to be on the test.

If at all possible, don't make the students memorize formulas.

In the order of topics, the main difference from a typical calculus textbook is that sequences come before everything else. The reasons are the following:

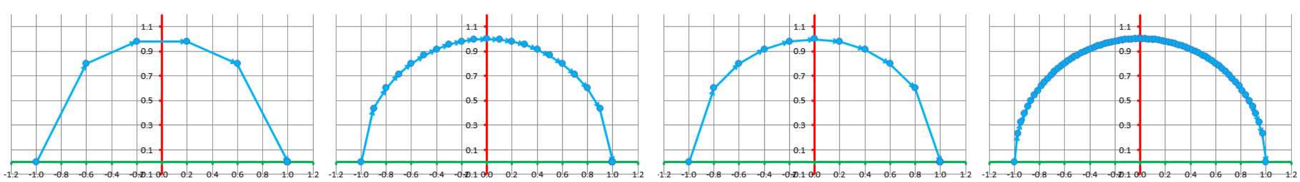
- Sequences are the simplest kind of functions.
- Limits of sequences are simpler than limits of general functions (including the ones at infinity).
- The sigma notation, the Riemann sums, and the Riemann integral make more sense to a student with a solid background in sequences.
- A quick transition from sequences to series often leads to confusion between the two.
- Sequences are needed for modeling, which should start as early as possible.

From the discrete to the continuous

It's no secret that a vast majority of calculus students will never use what they have learned. Poor career choices aside, a former calculus student is often unable to recognize the mathematics that is supposed to surround him. Why does this happen?

Calculus is the science of change. From the very beginning, its peculiar challenge has been to study and measure *continuous* change: curves and motion along curves. These curves and this motion are represented by *formulas*. Skillful manipulation of those formulas is what solves calculus problems. For over 300 years, this approach has been extremely successful in sciences and engineering. The successes are well-known: projectile motion, planetary motion, flow of liquids, heat transfer, wave propagation, etc. Teaching calculus follows this approach: An overwhelming majority of what the student does is manipulation of formulas on a piece of paper. But this means that all the problems the student faces were (or could have been) solved in the 18th or 19th centuries!

This isn't good enough anymore. What has changed since then? The computers have appeared, of course, and computers don't manipulate formulas. They don't help with solving – in the traditional sense of the word – those problems from the past centuries. Instead of *continuous*, computers excel at handling *incremental* processes, and instead of formulas they are great at managing discrete (digital) data. To utilize these advantages, scientists "discretize" the results of calculus and create algorithms that manipulate the digital data. The solutions are approximate but the applicability is unlimited. Since the 20th century, this approach has been extremely successful in sciences and engineering: aerodynamics (airplane and car design), sound and image processing, space exploration, structure of the atom and the universe, etc. The approach is also circuitous: Every concept in calculus *starts* – often implicitly – as a discrete approximation of a continuous phenomenon!



Calculus is the science of change, *both* incremental and continuous. The former part – the so-called discrete calculus – may be seen as the study of incremental phenomena and the quantities *indivisible* by their very nature: people, animals, and other organisms, moments of time, locations of space, particles, some commodities, digital images and other man-made data, etc. With the help of the calculus machinery called "limits", we invariably choose to transition to the continuous part of calculus, especially when we face continuous phenomena and the quantities *infinitely divisible* either by their nature or by assumption: time, space, mass, temperature, money, some commodities, etc. Calculus produces definitive results and absolute accuracy – but only for problems amenable to its methods! In the classroom, the problems are simplified until they become manageable; otherwise, we circle back to the discrete methods in search of approximations.

Within a typical calculus course, the student simply never gets to complete the "circle"! Later on, the graduate is likely to think of calculus only when he sees formulas and rarely when he sees numerical data.

In this book, every concept of calculus is first introduced in its discrete, "pre-limit", incarnation – elsewhere typically hidden inside proofs – and then used for modeling and applications well before its continuous counterpart emerges. The properties of the former are discovered first and then the matching properties of the latter are found by making the increment smaller and smaller, at the *limit*:

$$\boxed{\begin{array}{c}\text{discrete}\\ \text{calculus}\end{array}} \xrightarrow{\quad \Delta x\to 0 \quad} \boxed{\begin{array}{c}\text{continuous}\\ \text{calculus}\end{array}}$$

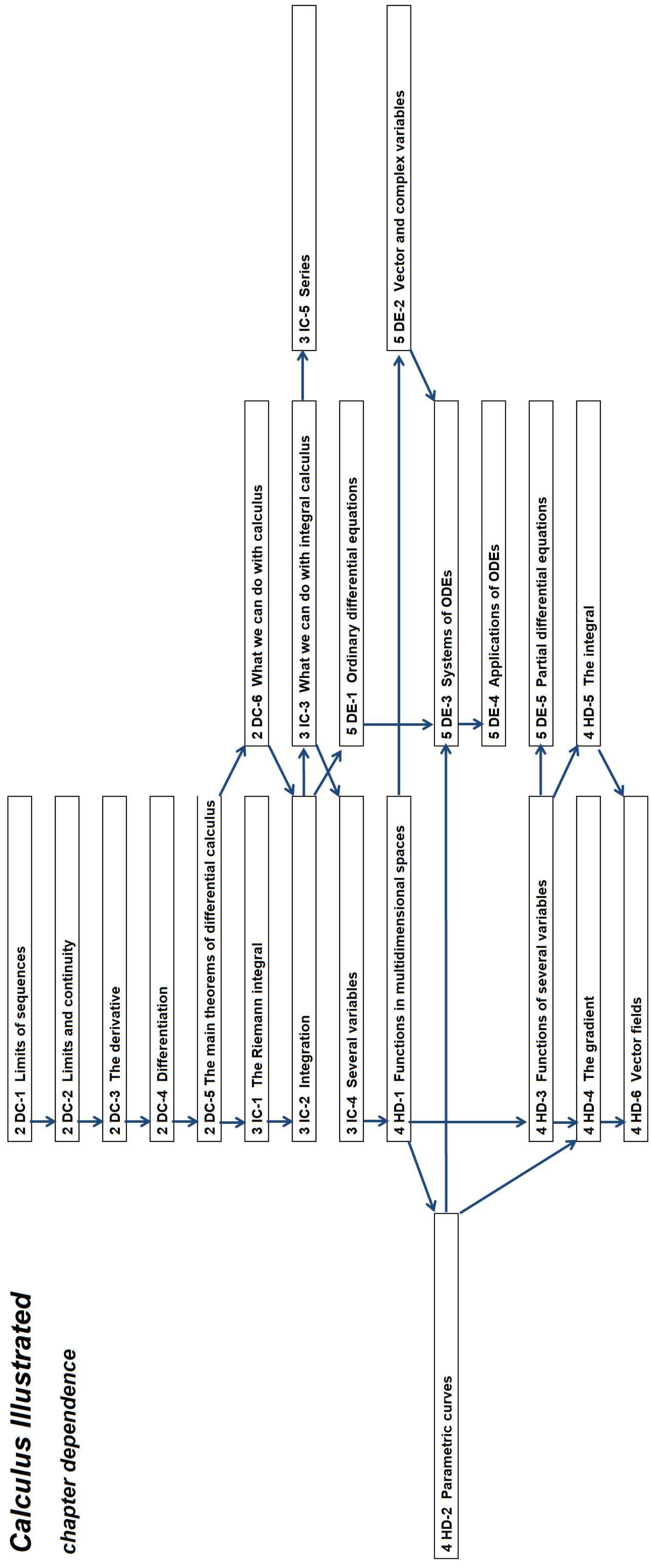
The volume and chapter references for *Calculus Illustrated*

This book is a part of the series *Calculus Illustrated*. The series covers the standard material of the undergraduate calculus with a substantial review of precalculus and a preview of elementary ordinary and partial differential equations. Below is the list of the books of the series, their chapters, and the way the present book (parenthetically) references them.

	■ **Calculus Illustrated. Volume 1: Precalculus**
1 PC-1	Calculus of sequences
1 PC-2	Sets and functions
1 PC-3	Compositions of functions
1 PC-4	Classes of functions
1 PC-5	Algebra and geometry
	■ **Calculus Illustrated. Volume 2: Differential Calculus**
2 DC-1	Limits of sequences
2 DC-2	Limits and continuity
2 DC-3	The derivative
2 DC-4	Differentiation
2 DC-5	The main theorems of differential calculus
2 DC-6	What we can do with calculus
	■ **Calculus Illustrated. Volume 3: Integral Calculus**
3 IC-1	The Riemann integral
3 IC-2	Integration
3 IC-3	What we can so with integral calculus
3 IC-4	Several variables
3 IC-5	Series
	■ **Calculus Illustrated. Volume 4: Calculus in Higher Dimensions**
4 HD-1	Functions in multidimensional spaces
4 HD-2	Parametric curves
4 HD-3	Functions of several variables
4 HD-4	The gradient
4 HD-5	The integral
4 HD-6	Vector fields
	■ **Calculus Illustrated. Volume 5: Differential Equations**
5 DE-1	Ordinary differential equations
5 DE-2	Vector and complex variables
5 DE-3	Systems of ODEs
5 DE-4	Applications of ODEs
5 DE-5	Partial differential equations

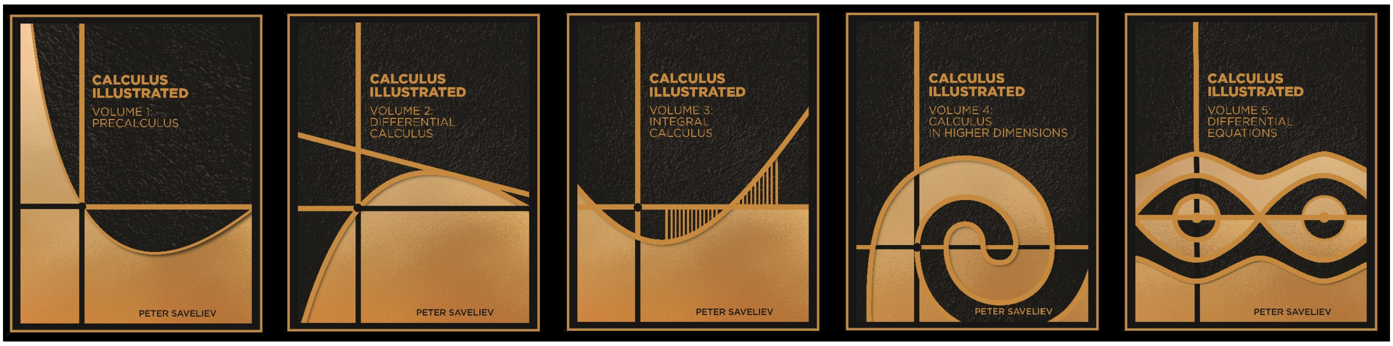
Each volume can be read independently.

A possible sequence of chapters is presented below. An arrow from A to B means that chapter B shouldn't be read before chapter A.



About the author

Peter Saveliev is a professor of mathematics at Marshall University, Huntington, West Virginia, USA. After a Ph.D. from the University of Illinois at Urbana-Champaign, he devoted the next 20 years to teaching mathematics. Peter is the author of a graduate textbook *Topology Illustrated* published in 2016. He has also been involved in research in algebraic topology and several other fields. His non-academic projects have been: digital image analysis, automated fingerprint identification, and image matching for missile navigation/guidance.



Contents

Chapter 1: The Riemann integral

Contents

1.1. The Area Problem

There are two main ways to enter calculus:

- Study of motion: back and forth between positions and velocities.
- Study of curved shapes: back and forth between tangents and... what? The areas.

We know that the *area of a circle* of radius r is supposed to be $A = \pi r^2$.

Example 1.1.1: area of circle

Let's review how we can confirm the formula with nothing but a spreadsheet.

First, we plot the graph

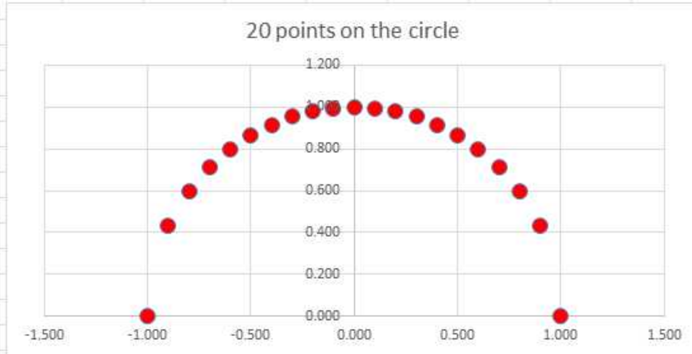
$$y = \sqrt{1 - x^2}\ ,$$

by letting the values of x run from -1 to 1 every $h = .1$ and finding the values of y with the spreadsheet formula:

```
=SQRT(1-RC[-2]^2)
```

We plot these 20 points; the result is a half-circle:

x	y
-1.000	0.000
-0.900	0.436
-0.800	0.600
-0.700	0.714
-0.600	0.800
-0.500	0.866
-0.400	0.917
-0.300	0.954
-0.200	0.980
-0.100	0.995
0.000	1.000
0.100	0.995
0.200	0.980
0.300	0.954
0.400	0.917
0.500	0.866
0.600	0.800
0.700	0.714
0.800	0.600

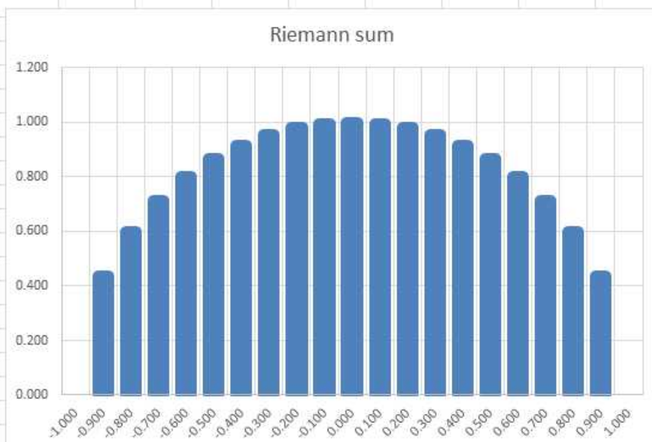


We next cover, as best we can, this half-circle with vertical bars that stand on the interval $[-1,1]$. We re-use the data:

- The bases of the bars are our intervals in the x-axis.
- The heights are the values of y.

To see the bars, we simply change the type of the chart plotted by the spreadsheet:

n=	20	intervals			
h=	0.10		Area=	1.552	1.571
	x	widths	heights	areas	
	-1.000		0.000		
	-0.900	0.100	0.436	0.044	
	-0.800	0.100	0.600	0.060	
	-0.700	0.100	0.714	0.071	
	-0.600	0.100	0.800	0.080	
	-0.500	0.100	0.866	0.087	
	-0.400	0.100	0.917	0.092	
	-0.300	0.100	0.954	0.095	
	-0.200	0.100	0.980	0.098	
	-0.100	0.100	0.995	0.099	
	0.000	0.100	1.000	0.100	
	0.100	0.100	0.995	0.099	
	0.200	0.100	0.980	0.098	
	0.300	0.100	0.954	0.095	
	0.400	0.100	0.917	0.092	
	0.500	0.100	0.866	0.087	
	0.600	0.100	0.800	0.080	
	0.700	0.100	0.714	0.071	



Next, the area of the circle is approximated by the *sum* of the areas of the bars. To let the spreadsheet do the work for us:

- Multiply the heights by the (constant) widths in the next column.
- Add them up at the top cell (yellow).

We use the formulas for *areas*:

```
=RC[-1]*(RC[-2]-R[-1]C[-2])
```

The result produced is the following:

$$\text{Approximate area of the semicircle} = 1.552\,.$$

It is close to the theoretical result obtained later in this chapter:

$$\text{Exact area of the semicircle} = \pi/2 \approx 1.571\,.$$

In summary, the area A is the sum of the sequence:

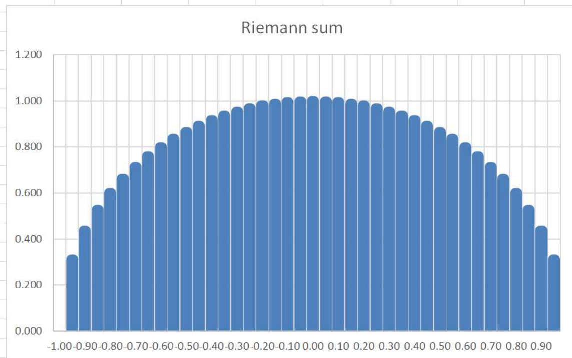
$$a_n = \sqrt{1-x_n^2}\cdot 0.1, \text{ where } \ x_n = -1.0\,,\ -0.9\,,\ -0.8\,,\ \ldots\,,\ 0.8\ ,\ 0.9\,,\ 1.0\,.$$

In other words:

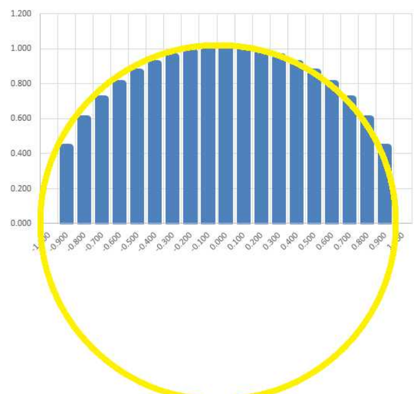
$$A = \sum_{n=1}^{20} 0.1\cdot\sqrt{1-x_n^2}\,.$$

There is more! The approach we have used many times in calculus (Chapter 2DC-3) is to *improve* our approximation by making the intervals smaller and smaller. Redoing the computation with 40 intervals gives a better approximation, 1.564:

n=	40	intervals			
h=	0.05		Area=	1.564	1.571
	x	widths	heights	areas	
	-1.000		0.000		
	-0.950	0.050	0.312	0.016	
	-0.900	0.050	0.436	0.022	
	-0.850	0.050	0.527	0.026	
	-0.800	0.050	0.600	0.030	
	-0.750	0.050	0.661	0.033	
	-0.700	0.050	0.714	0.036	
	-0.650	0.050	0.760	0.038	
	-0.600	0.050	0.800	0.040	
	-0.550	0.050	0.835	0.042	
	-0.500	0.050	0.866	0.043	
	-0.450	0.050	0.893	0.045	
	-0.400	0.050	0.917	0.046	
	-0.350	0.050	0.937	0.047	
	-0.300	0.050	0.954	0.048	
	-0.250	0.050	0.968	0.048	
	-0.200	0.050	0.980	0.049	
	-0.150	0.050	0.989	0.049	
	-0.100	0.050	0.995	0.050	



The quality of an approximation is seen as the size of the parts of the bars sticking out of the circle as well as the parts of the inside of the circle not covered by the bars:



Nothing stops us from improving this approximation further and further with larger and larger values of n.

Exercise 1.1.2

Approximate the area of the circle of radius 1 within 0.0001.

We have showed that indeed the area of the circle is close to what's expected. But the real question is:

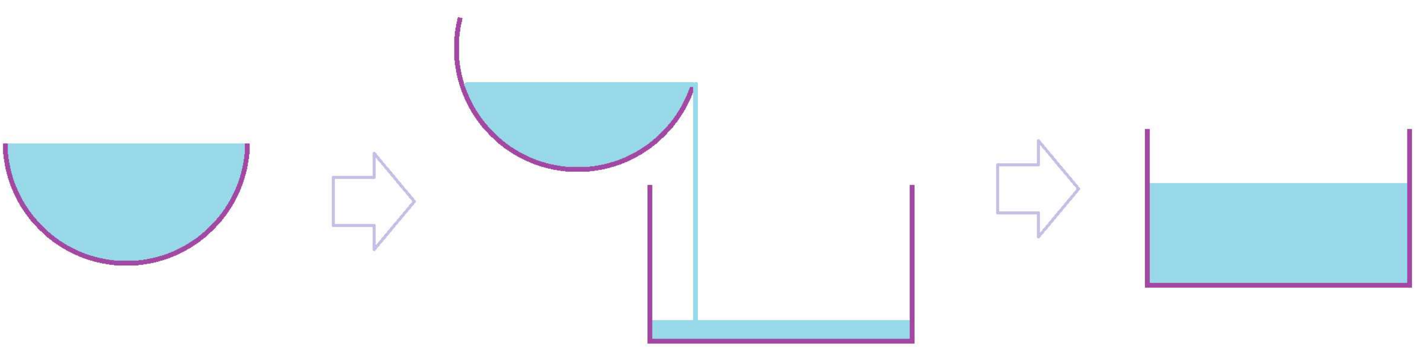
- What *is* the area?

Do we even understand what it is?

Warning!

The area is a number.

The intuition is to speak of the amount of material held by the curve. For example, we fill a semicircular bucket with water and the pour it into a rectangular one so that we can measure the contents:

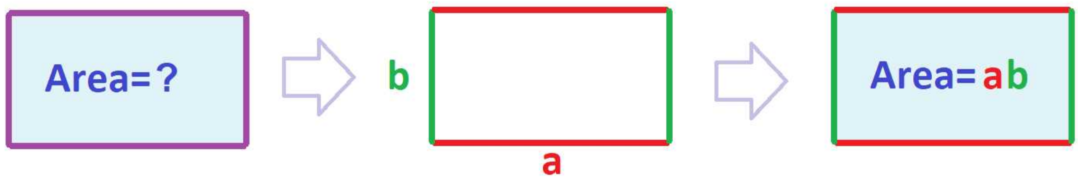


Exercise 1.1.3

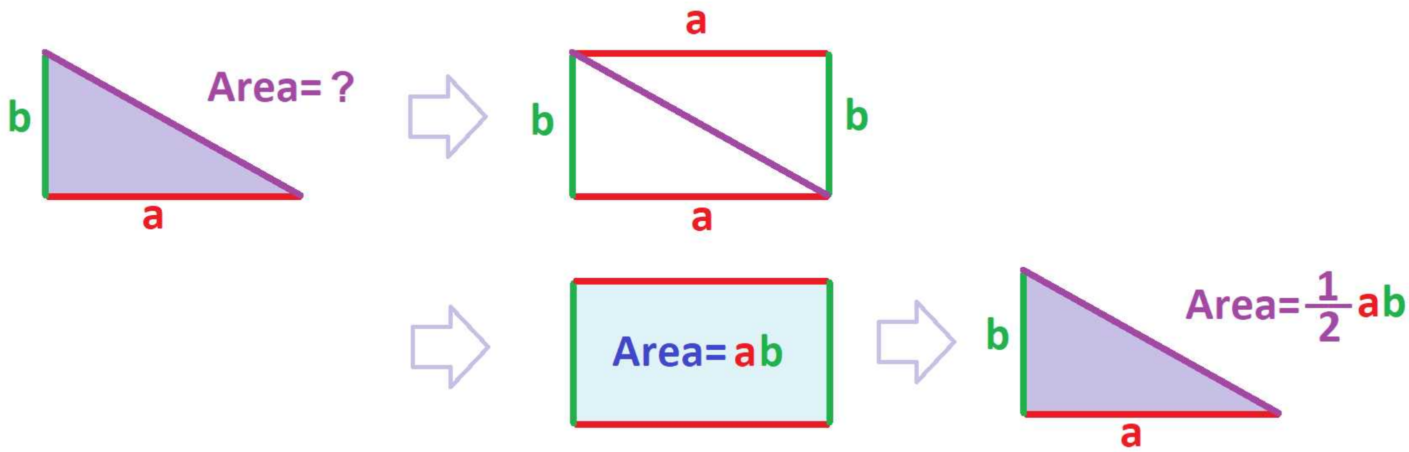
Confirm that the semicircular bucket and the rectangular one contain the same amount of water.

How do we make mathematical sense of this?

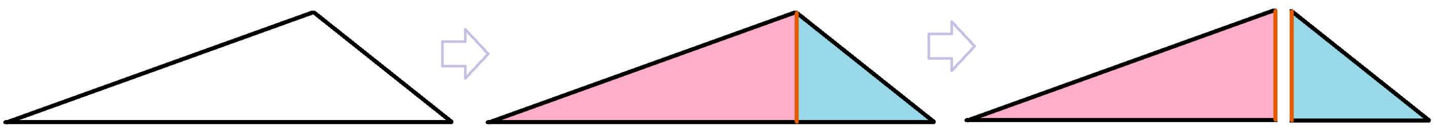
One thing we do know. The area of a rectangle $a \times b$ is ab.



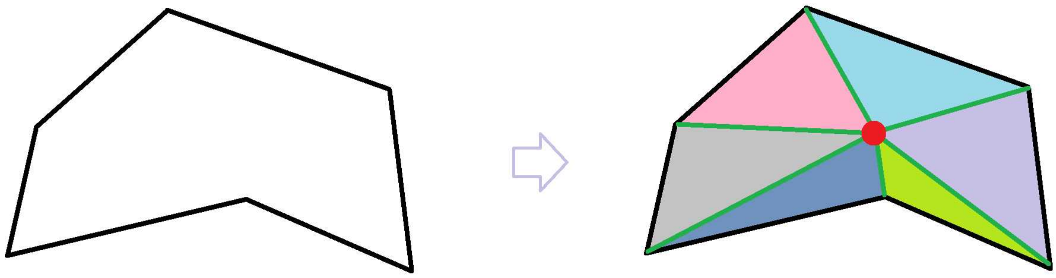
Furthermore, any right triangle is simply a half of a diagonally cut rectangle:



We can also cut any triangle into a pair of right triangles:



Finally, any polygon can be cut into triangles:

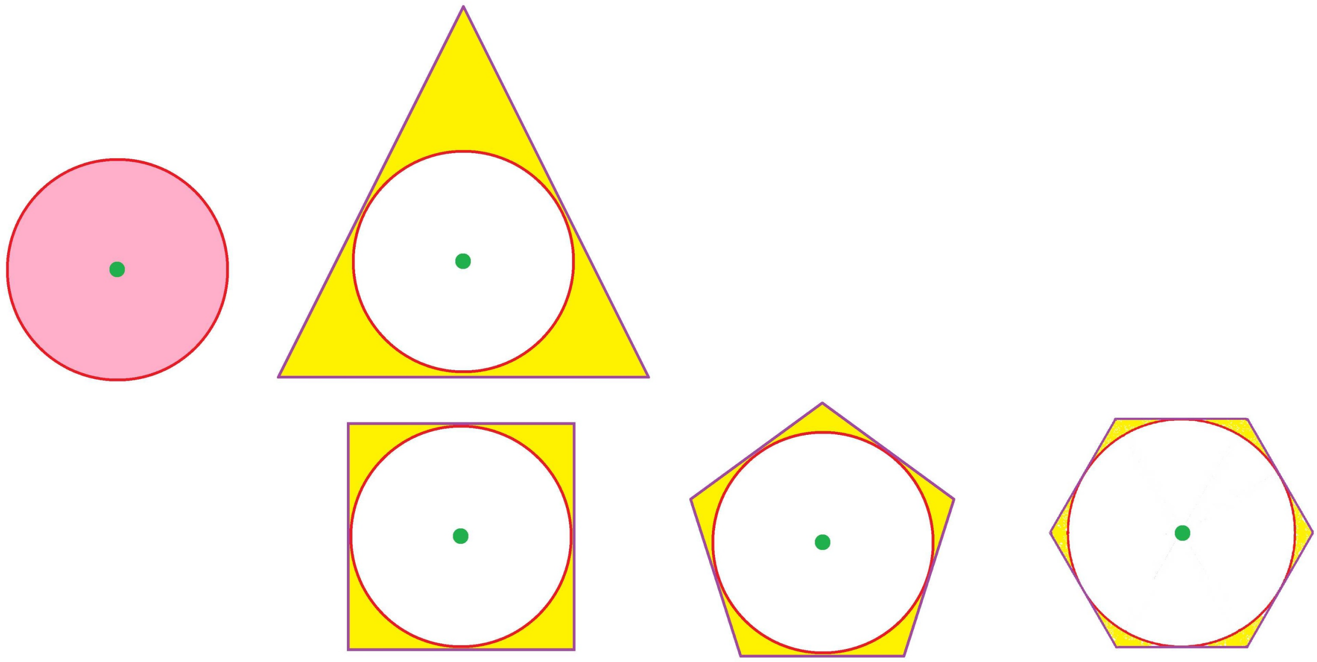


So, we can find – and we understand – the areas of all polygons. They are geometric objects with *straight* edges.

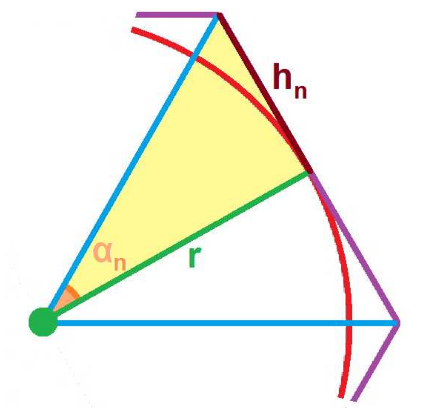
But what are the *areas of curved* objects?

Example 1.1.4: circle

Let's review the ancient Greeks' approach to understanding and computing the area of a circle. They approximated the circle with regular polygons: equal sides and angles. We put such polygons around the circle so that it touches them from the inside ("circumscribing" polygons):



For each $n = 3, 4, 5, 6, ...$, the following is carried out. We split each such polygon with n sides into $2n$ right triangles and compute their dimensions:



We think of π here as the angle measure (possibly unknown) of a half of the full turn. The full turn is 2π, and it is cut into $2n$ angles:

$$\alpha_n = \frac{2\pi}{2n} = \frac{\pi}{n}\,.$$

The side that touches the circle is its radius and the opposite side is $r \tan \frac{\pi}{n}$. Therefore, the area of the triangle is

$$a_n = \frac{1}{2} \cdot r \cdot r \tan \tfrac{\pi}{n} = \frac{r^2}{2} \tan \tfrac{\pi}{n}\,.$$

We know now the area of the whole polygon:

$$A_n = a_n \cdot 2n = \frac{r^2}{2} \tan \tfrac{\pi}{n} \cdot 2n\,.$$

We can examine the data:

n	3	4	5	6	7	8	9	10	11	12	13	14	15	...
A_n	5.196	4.000	3.633	3.464	3.371	3.314	3.276	3.249	3.230	3.215	3.204	3.195	3.188	...

The numbers seem to *converge*!

Indeed, the sequence A_n is both monotone and bounded and, therefore, convergent. Its limit is the *meaning* of the area (and of π).

Let's compute the limit:

$$\begin{aligned} A_n &= \frac{r^2}{2}\tan\tfrac{\pi}{n}\cdot 2n \\ &= \pi r^2 \cdot \frac{\sin\frac{\pi}{n}}{\frac{\pi}{n}} \cdot \frac{1}{\cos\frac{\pi}{n}} \\ &\quad || \qquad\ \ \downarrow \qquad \downarrow \qquad \text{as } n\to\infty \\ &\quad \pi r^2 \quad \cdot 1 \qquad \cdot 1 \\ &= \pi r^2 . \end{aligned}$$

We rearrange the terms in order to use...

one of the famous trig limits (Chapter 2DC-1).

In order to fully justify the result, the Greeks also put such polygons around the circle so that it touches them from the *outside* ("inscribing" polygons). Unfortunately, the methods cannot be easily applied to, say, parabolas. That is why we will seek a more general approach.

Exercise 1.1.5

Explain in detail why A_n converges.

Exercise 1.1.6

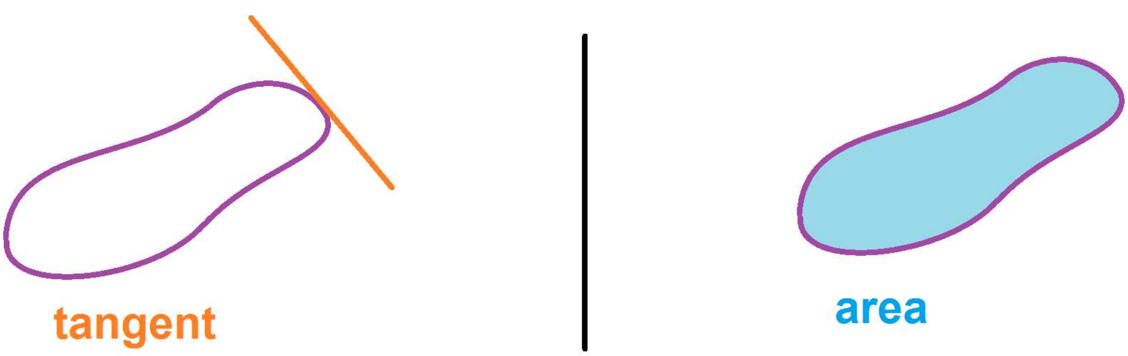
Explain the other limit in our computation.

Exercise 1.1.7

Carry out this construction for the inscribed polygons.

Let's compare these two seemingly unrelated problems and how they are solved:

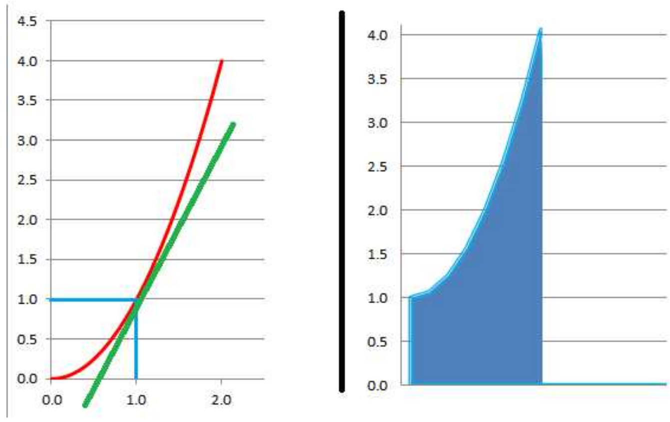
The Tangent Problem	**The Area Problem**



Geometry: For a given curve, find the following:

the line touching the curve at a point	the area enclosed by the curve

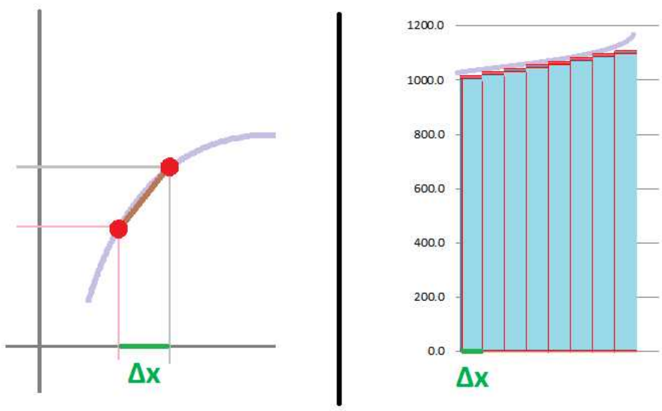
The problems are easy when there is no actual curving: straight lines. Furthermore, we have solved these problems for a specific curve: the circle. The further progress depends on the development of algebra, the Cartesian coordinate system, and the idea of function (Chapters 1PC-2, 1PC-3, and 1PC-4).



Motion: The two problems are interpreted as follows:

The slope of the tangent to the position function is the velocity at that moment.	The area under the graph of the velocity function is the displacement.

The further development utilizes dividing the domain of the function into smaller pieces, Δx long, sampling the function, and approximating the function by means of straight lines.



For many practical applications, this stage is sufficient. In idealized situations, we can do more.

Calculus: The limits of these approximations, as $\Delta x \to 0$, are the following:

the derivative	the integral

In this chapter, we will pursue the plan for the latter problem as outlined in the right column.

1.2. Differences and sums

We turn to *motion* now. Our starting point is the formula:

$$\boxed{\text{distance} = \text{speed} \times \text{time}}$$

Recall two familiar problems (Chapter 2DC-1).

Problem: Imagine that our speedometer is broken and we need to find a way to estimate how fast we are driving.

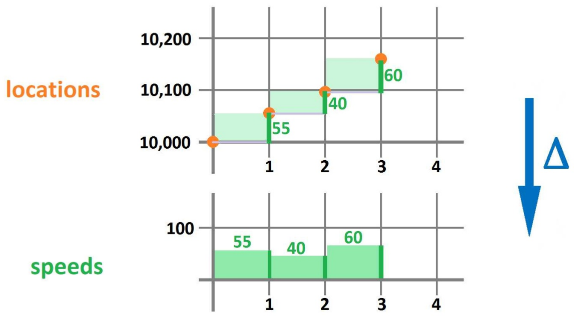
We look at the odometer *several* times during the trip and record the mileage on a piece of paper:

1. initial reading: $10,000$ miles
2. after the first hour: $10,055$ miles
3. after the second hour: $10,095$ miles
4. after the third hour: $10,155$ miles
5. etc.

That's a sequence. We now use the *difference of the sequence* to solve the problem:

1. distance covered during the first hour: $10,055 - 10,000 = 55$ miles
2. distance covered during the second hour: $10,095 - 10,055 = 40$ miles
3. distance covered during the third hour: $10,155 - 10,095 = 60$ miles
4. etc.

That's another sequence. We see below how these new numbers appear as the heights of the steps of our last plot (top):



As you can see, we illustrate the new data in such a way as to suggest that the speed remains *constant* during each of these hour-long periods.

In this chapter, we will concentrate on the flip side of the last problem.

Problem: Imagine that it is the odometer that is broken and find a way to estimate how far we will have gone.

Of course, we look at the speedometer *several* times – say, every hour – during the trip and record its readings on a piece of paper:

1. during the first hour: 35 miles an hour
2. during the second hour: 65 miles an hour
3. during the third hour: 50 miles an hour
4. etc.

That's a sequence. What does this tell us about our location? Nothing, without algebra! Fortunately, we can just use the same formula as before.

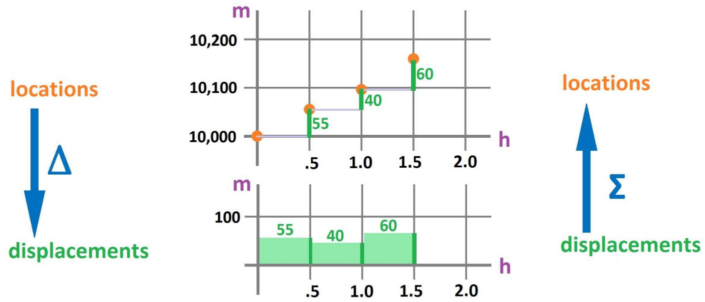
In contrast to the former problem, we need another bit of information. We must know the *starting point* of our trip, say, the 100-mile mark. The time interval was chosen to be 1 hour so that we need only to *add*, and keep adding, the speed at which – we assume – we drove during each of these one-hour periods:

1. the initial location: 100-mile mark
2. the location after the first hour: $100 + 35 = 135$-mile mark
3. the location after the second hour: $135 + 65 = 200$-mile mark
4. the location after the third hour: $200 + 50 = 250$-mile mark

5. etc.

That's another sequence. It is called the *sum of the sequence*.

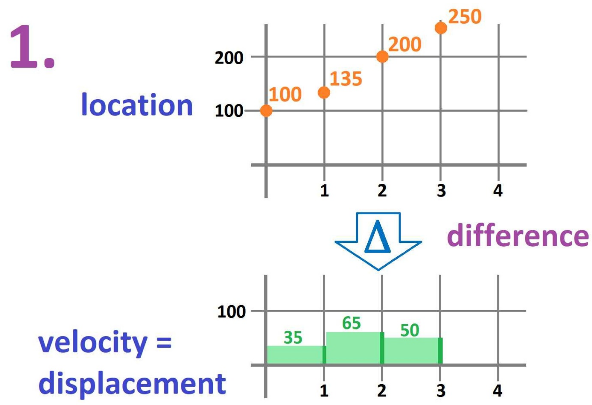
In order to illustrate this algebra, we use the speeds as the heights of the consecutive steps of the staircase:

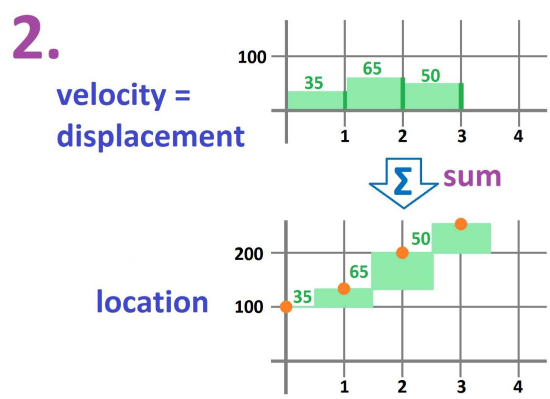


Then the new numbers show how high we have to climb in our last plot.

The problem is solved! We have established that we have progressed through the roughly 135-, 200-, and 250-mile marks during this time.

In summary, when the intervals of time are units, we can go from locations to velocities and back with the two simple operations:





This is the summary:

1. If each term of a sequence represents a location, the pairwise differences will give you the *velocities*. A broken speedometer is substituted with an odometer and a watch. This study is appeared in Volume 2, Chapter 2DC-3.
2. If each term of a sequence represents a velocity, their sum up to that point will give you the *location*. A broken odometer is substituted with a speedometer and a watch. This study will appear in this chapter.

In the abstract, the pairwise differences represent the *change* within the sequence, from each of its terms to the next:

Definition 1.2.1: sequence of differences

For a sequence a_n, its *sequence of differences*, or simply the difference, is a new sequence, say d_n, defined for each n by the following:

$$d_n = a_{n+1} - a_n \,.$$

It is denoted as follows:

$$\boxed{\Delta a_n = a_{n+1} - a_n}$$

In this section, we start on the path of development of an idea that culminates with the second fundamental concept of calculus to be seen later in this chapter.

The sum represents the totality of the "beginning" of a sequence, found by adding each of its terms to the next, up to that point.

Example 1.2.2: sequences given by lists

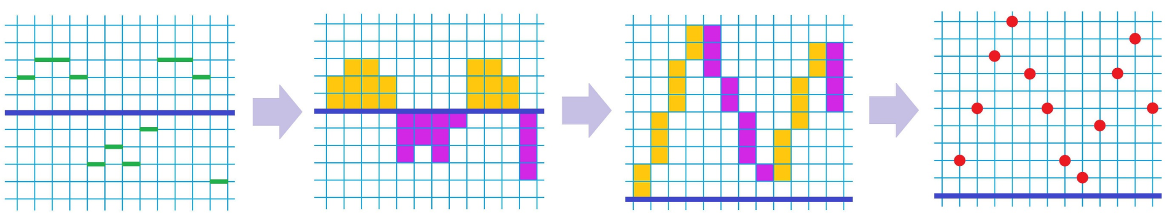
We just add the current term to what we have accumulated so far:

$$\begin{array}{rcccccccccccccccccc}
\text{sequence:} & 2 & & 4 & & & & 7 & & & & 1 & & & & -1 & & & \dots \\
 & \downarrow & & \downarrow & & & & \downarrow & & & & \downarrow & & & & \downarrow & & & \dots \\
\text{sums:} & 2 & & & & & & & & & & & & & & & & & \\
 & 2 & + & 4 & = & 6 & & & & & & & & & & & & & \\
 & & & & & 6 & + & 7 & = & 13 & & & & & & & & & \\
 & & & & & & & & & 13 & + & 1 & = & 14 & & & & & \\
 & & & & & & & & & & & & & 14 & + & (-1) & = & 13 & \\
 & \downarrow & & & & \downarrow & & & & \downarrow & & & & \downarrow & & & & \downarrow & \dots \\
\text{new sequence:} & 2 & & & & 6 & & & & 13 & & & & 14 & & & & 13 & \dots
\end{array}$$

We have a new list!

Example 1.2.3: sequences given by graphs

We treat the graph of a sequence as if made of bars and then just stack up these bars on top of each other one by one:



These stacked bars – or rather the process of stacking – make a new sequence.

Definition 1.2.4: sequence of sums

For a sequence a_n, its *sequence of sums*, or simply the sum, is a new sequence s_n defined and denoted for each $n \ge m$ within the domain of a_n by the following (recursive) formula:

$$\boxed{s_m = 0, \quad s_{n+1} = s_n + a_{n+1}}$$

In other words,

$$\boxed{s_n = a_m + a_{m+1} + ... + a_n}$$

The following is an abbreviated way to present summation of a sequence.

Sigma notation for summation

$$s_n = a_m + a_{m+1} + ... + a_n = \sum_{k=m}^{n} a_k$$

Warning!

Using either "..." and "$\sum$" might obscure the recursive nature of the computation.

Let's take a closer look at the new notation. The first choice of how to represent the sum of a segment – from m to n – of a sequence a_n is this:

$$\underbrace{a_m}_{\text{step 1}} \underbrace{+a_{m+1}}_{\text{step 2}} + \ldots \underbrace{+a_k}_{\text{step } k} + \ldots \underbrace{+a_n}_{\text{step } n-m}$$

This notation reflects the recursive nature of the process but it can also be repetitive and cumbersome. The new notation is meant to make it more compact. The idea is to introduce an "internal variable" k as follows:

Sigma notation

$$\sum_{k=0}^{3}\left(k^2+k\right)=20 \quad \longrightarrow \quad \overbrace{\sum_{k=0}^{3}}^{\text{beginning and end values for } k} \Big(\underbrace{k^2+k}_{\text{a specific sequence of } k}\Big) = \underbrace{20}_{\text{a specific number}}$$

Warning!

It would also make sense to have "$k=3$" above the sigma:

$$\sum_{k=0}^{k=3}\left(k^2+k\right).$$

Here the Greek letter Σ stands for the letter S meaning "sum".

Example 1.2.5: expanding from sigma notation

The computation above is expanded here:

$$\sum_{k=0}^{3}\left(k^2+k\right) = \begin{array}{c|lll} k & k^2+k & & \\ \hline 0 & 0^2+0 & =0 & + \\ 1 & 1^2+1 & =2 & + \\ 2 & 2^2+2 & =6 & + \\ 3 & 3^2+3 & =12 & \\ \hline & & =20 & \end{array}$$

Exercise 1.2.6: contracting to sigma notation

How will the sum change if we replace $k=0$ with $k=1$, or $k=-1$? What if we replace 3 at the top with 4?

Example 1.2.7: contracting summation

This is how we *contract* the summation:

$$1^2+2^2+3^2+\ldots+17^2=\sum_{k=1}^{17}k^2\,.$$

This is only possible if we find the nth-term formula for the sequence; in this case, $a_k=k^2$.

And this is how we *expand* back from this compact notation, by plugging the values of $k=1,2,\ldots,17$

into the formula:

$$\sum_{k=1}^{17} k^2 = \underbrace{1^2}_{k=1} + \underbrace{2^2}_{k=2} + \underbrace{3^2}_{k=3} + ... + \underbrace{17^2}_{k=17} .$$

Similarly, we have:

$$1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + ... + \frac{1}{2^{10}} = \sum_{k=0}^{10} \frac{1}{2^k} .$$

Exercise 1.2.8

Confirm that we can start at any other initial index if we just modify the formula:

$$1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + ... + \frac{1}{2^{10}} = \sum_{k=?}^{?} \frac{1}{2^{k-1}} = \sum_{k=?}^{?} \frac{1}{2^{k-2}} = ...$$

Exercise 1.2.9

Contract this summation:

$$1 + \frac{1}{3} + \frac{1}{9} + \frac{1}{27} = ?$$

Exercise 1.2.10

Expand this summation:

$$\sum_{k=0}^{4} (k/2) = ?$$

Exercise 1.2.11

Rewrite using the sigma notation:

1. $1 + 3 + 5 + 7 + 9 + 11 + 13 + 15$
2. $0.9 + 0.99 + 0.999 + 0.9999$
3. $1/2 - 1/4 + 1/8 - 1/16$
4. $1 + 1/2 + 1/3 + 1/4 + ... + 1/n$
5. $1 + 1/2 + 1/4 + 1/8$
6. $2 + 3 + 5 + 7 + 11 + 13 + 17$
7. $1 - 4 + 9 - 16 + 25$

The notation applies to all sequences, both finite and infinite. For infinite sequences, recognized by "..." at the end, the sum sequence is also called "partial sums" as well as "series" (to be discussed in Chapter 4).

This is the recursive definition of the sequence of sums re-written with the sigma notation:

Sequence of sums

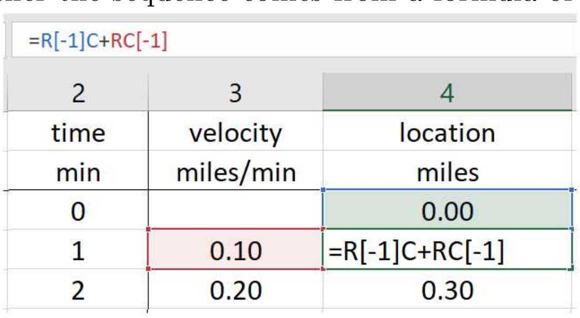
$$
\begin{array}{lccccccccccccc}
\text{a sequence:} & a_1 & & a_2 & & & a_3 & & & a_4 & & & \dots \\
 & \downarrow & & \downarrow & & & \downarrow & & & \downarrow & & & \dots \\
\text{its sums:} & a_1 & & & & & & & & & & & \\
 & a_1 & + & a_2 & = & s_2 & & & & & & & \\
 & & & & & s_2 & + & a_3 & = & s_3 & & & \\
 & & & & & & & & & s_3 & + & a_4 & = & s_4 \quad \dots \\
 & & & & & & & & & & & & & s_4 \quad \dots \\
 & \downarrow & & & & \downarrow & & & & \downarrow & & & & \downarrow \quad \dots \\
\text{the sequence of sums:} & s_1 & & & & s_2 & & & & s_3 & & & & s_4 \quad \dots \\
 & || & & & & || & & & & || & & & & || \quad \dots \\
\text{the sigma notation:} & \sum_{k=1}^{1} a_k & & & & \sum_{k=1}^{2} a_k & & & & \sum_{k=1}^{3} a_k & & & & \sum_{k=1}^{4} a_k \quad \dots
\end{array}
$$

Example 1.2.12: sums are displacements

We can use computers to speed up these computations. For example, one may have been recording one's velocities and now looking for the location. This is a formula for a spreadsheet (the locations):

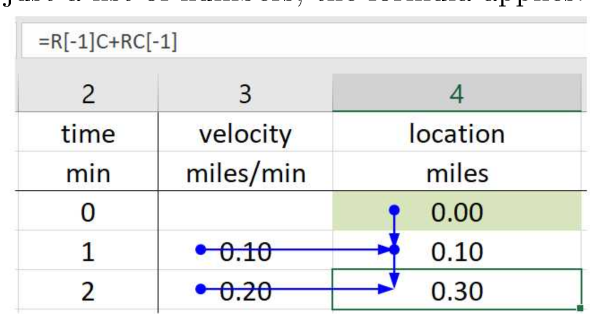
```
=R[-1]C+RC[-1]
```

Whether the sequence comes from a formula or it's just a list of numbers, the formula applies:



=R[-1]C+RC[-1]

2	3	4
time	velocity	location
min	miles/min	miles
0		0.00
1	0.10	=R[-1]C+RC[-1]
2	0.20	0.30



=R[-1]C+RC[-1]

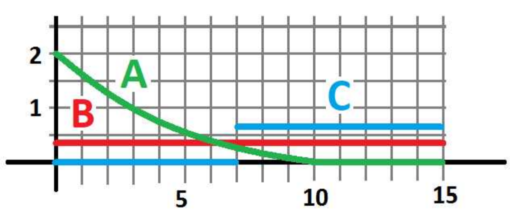
2	3	4
time	velocity	location
min	miles/min	miles
0		0.00
1	0.10	0.10
2	0.20	0.30

As a result, a curve has produced a new curve:

a_n	Σa_n
	0.0
9.0	9.0
23.0	32.0
31.0	63.0
33.0	96.0
29.0	125.0
19.0	144.0
3.0	147.0
-19.0	128.0
-47.0	81.0
-81.0	0.0
-121.0	-121.0

Example 1.2.13: three runners, continued

The graph shows the velocities of three runners in terms of time, n:

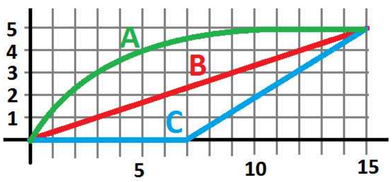


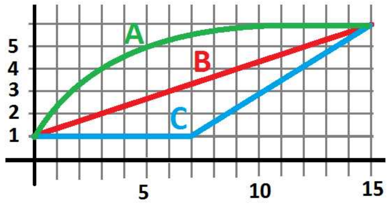
It's easy to describe *how* they are moving:

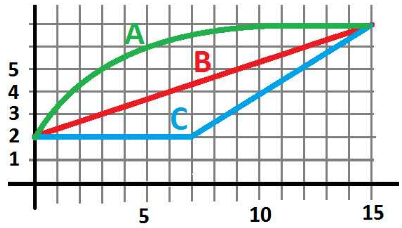
- A starts fast and then slows down.
- B maintains the same speed.
- C starts late and then runs fast.

But *where* are they, at every moment? A simple examination of the first graph shows that B and C will arrive at the finish line at the same time. To say that about A would require more subtle analysis.

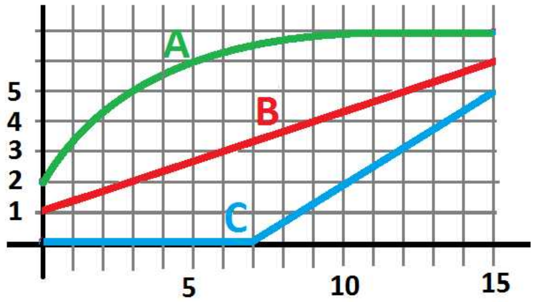
There are several possible answers:







Which one is the right one depends on the starting point. Furthermore, if the requirement that they all start at the same location is lifted, the result will be different, for example:



Exercise 1.2.14

Suggest other graphs that match the description above.

Exercise 1.2.15

Plot the location and the velocity for the following trip: "I drove slowly, gradually sped up, stopped for a very short moment; started again but in the opposite direction, quickly accelerated, and from that point maintained the speed." Make up your own story and repeat the task.

Exercise 1.2.16

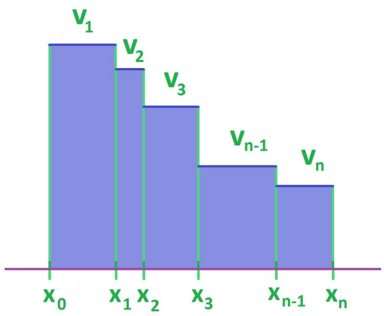
Draw a curve on a piece of paper, imagine that it represents your velocity, and then sketch what your locations would look like. Repeat.

How do we deal with motion when the time moments aren't integers? What is the *displacement* then?

Suppose

- x_n is the sequence of moments of time, and
- v_n is the corresponding sequence of velocities.

They are illustrated as follows:



The velocities are the heights of the rectangles. The lengths of the intervals are the lengths of the bases of the rectangles:

$$\Delta x_n = x_n - x_{n-1}\,.$$

As we know, the displacement during the time interval $[x_{n-1}, x_n]$ is the product of v_n and the difference of x_n:

$$\text{displacement } = v_n \cdot \Delta x_n\,.$$

Then the displacements form a sequence:

$$\text{displacements: } v_1 \cdot \Delta x_1,\ v_2 \cdot \Delta x_2,\ ...,\ v_n \cdot \Delta x_n,\ ...$$

Then the total displacement during the time interval $[x_0, x_n]$ is the sum of this sequence:

$$\text{total displacement } = v_1 \cdot \Delta x_1 + v_2 \cdot \Delta x_2 + ... + v_n \cdot \Delta x_n = \sum_{k=1}^{n} v_k \cdot \Delta x_k\,.$$

With this idea, we start to develop the mathematics of the back-and-forth interaction within these two pairs:

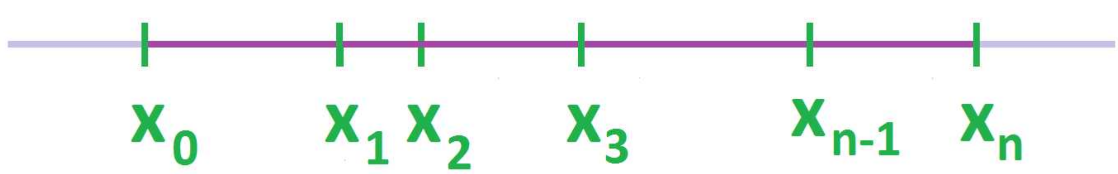
- positions and velocities
- tangents and areas

The general setup follows. There will be no restrictions whatsoever on the segments or the sample points in this construction.

First, we need a *partition* of an interval $[a, b]$. We choose a natural number n and then place $n + 1$ points on the interval:

$$a = x_0 \le x_1 \le x_2 \le ... \le x_{n-1} \le x_n = b\,.$$

It's an increasing sequence x_i:



As a result, the interval is partitioned into n smaller intervals of possibly different lengths:

$$[x_0, x_1],\ [x_1, x_2],\ ...,\ [x_{n-1}, x_n]\,.$$

Definition 1.2.17: partition of interval

A *partition of an interval* $[a, b]$ is its representation as the union of intervals that intersect only at the end-points:

$$[a, b] = [x_0, x_1] \cup [x_1, x_2] \cup ... \cup [x_{n-1}, x_n]\,.$$

These end-points will be called the *primary nodes* of the partition.

The difference of this sequence gives us the lengths of these intervals.

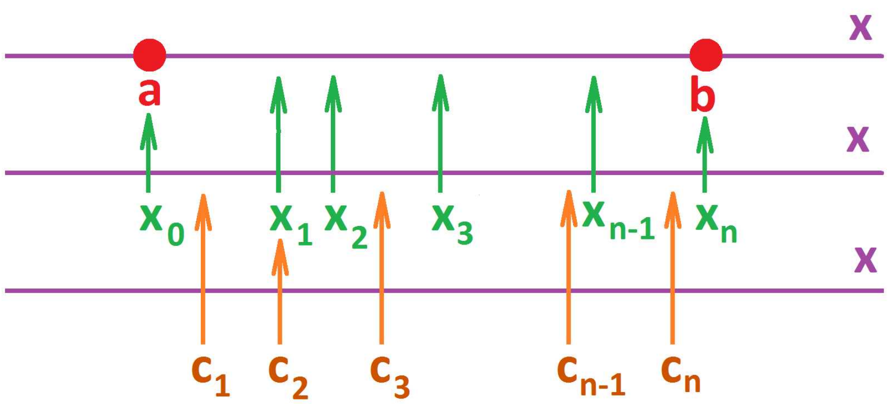
Definition 1.2.18: increments of partition

The *increments* of the partition are given by:

$$\Delta x_i = x_i - x_{i-1},\ i = 1, 2, ..., n\,.$$

In addition to the nodes, the primary nodes, we may also be given the *secondary nodes* in each interval of the partition:

$$c_1 \text{ in } [x_0, x_1],\ c_2 \text{ in } [x_1, x_2],\ ...,\ c_n \text{ in } [x_{n-1}, x_n]\,.$$



In summary:

Definition 1.2.19: augmented partition of interval

An *augmented partition*, or simply a partition, of an interval $[a, b]$ consists of two sequences:

1. *primary nodes* $a = x_0,\ x_1,\ x_2,\ ...,\ x_{n-1},\ x_n = b$;
2. *secondary nodes* $c_1,\ c_2,\ c_3,\ ...,\ c_{n-1},\ c_n$;

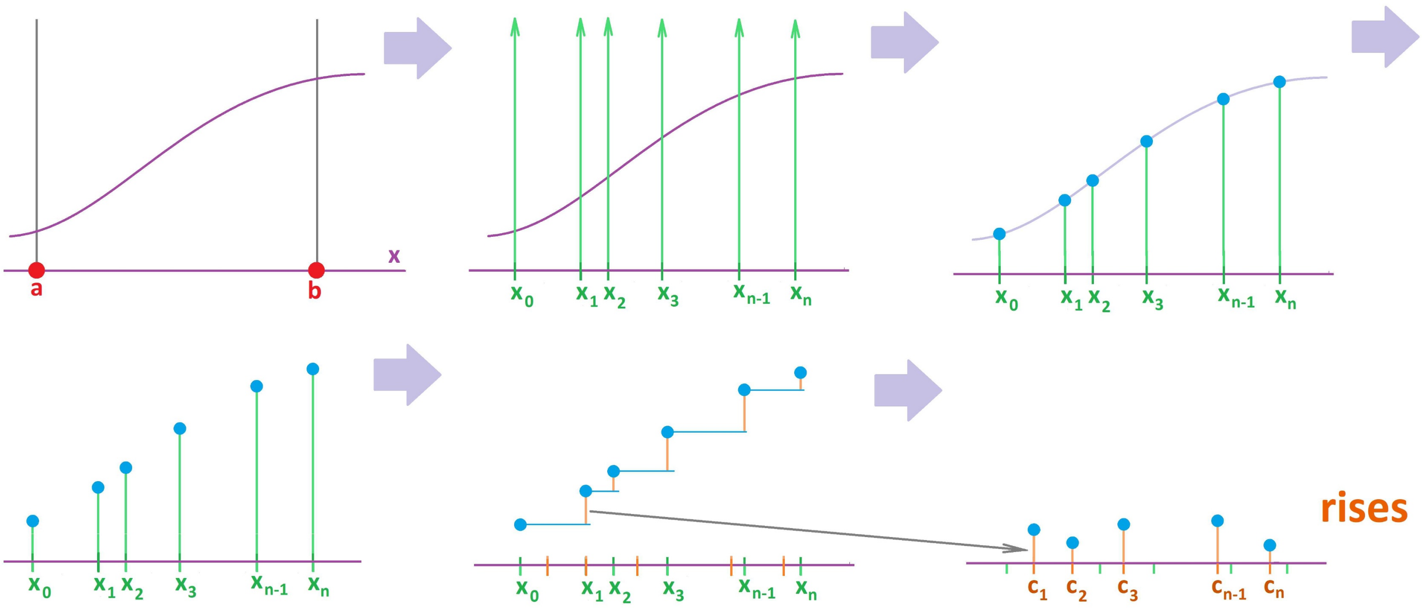
that satisfy these inequalities:

$$x_0 \le c_1 \le x_1 \le c_2 \le x_2 \le ... \le x_{n-1} \le c_n \le x_n\,.$$

Warning!

We can chose secondary nodes from the list of primary nodes because the inequalities are non-strict.

The construction of the differences of a function (i.e., the rises) is outlined below:



These are the stages that we see here:

1. a function
2. a partition and its primary nodes
3. sampling of the function at the primary nodes
4. removing the rest of the graph
5. plotting the differences (rises)
6. placing these differences at the secondary nodes

The result is a new function.

It is defined algebraically as follows:

Definition 1.2.20: difference of function

Suppose $y = f(x)$ is defined at the nodes $x_k,\ k = 0, 1, 2, ..., n$, of a partition. Then the *difference* of f is a function defined on each interval of the partition and, correspondingly, at every secondary node of the partition, and denoted, as follows:

$$\boxed{\Delta\, f\Big|_{[x_{k-1},x_k]} = \Delta f\,(c_k) = f(x_{k+1}) - f(x_k)}$$

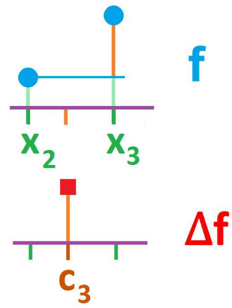
Let's remember that here the original function f is evaluated over the interval, $[x_{k-1}, x_k]$, while the new function Δf is defined at c_k:

$$\Delta\, f\Big|_{[x_{k-1},x_k]} = \Delta f\,(c_k)\,.$$

We can also use parentheses to separate functions from their inputs when it's not cumbersome:

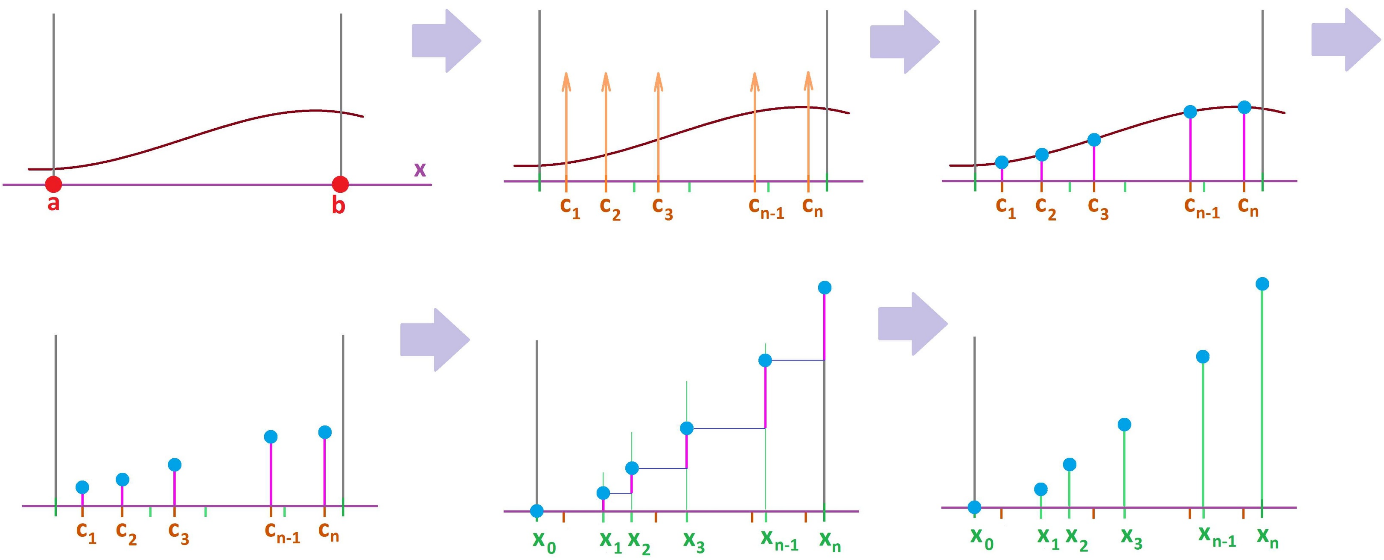
$$\Delta\left(f\Big|_{[x_{k-1},x_k]}\right) = (\Delta f)(c_k)\,.$$

The difference is computed one segment at a time:



Now we go in the opposite direction: from velocities to positions and from areas to tangents.

The construction of the differences of a function (i.e., the rises) is outlined below:



These are the stages that we see here:

1. a function
2. a partition and its secondary nodes
3. sampling of the function at the secondary nodes
4. removing the rest of the graph
5. putting these values on top of each other
6. placing these sums at the primary nodes

The result is a new function.

It is defined algebraically as follows:

Definition 1.2.21: sum of function

Suppose a function g defined at the secondary nodes of an augmented partition P of an interval $[a,b]$. The *sum* of g is the function h defined on the primary nodes as the sum of the sequence:

$$g(c_1),\ g(c_2),\ \dots,\ g(c_k),\ \dots$$

In other words, it is defined (recursively):

$$\boxed{h(x_0)=0,\quad h(x_m)=h(x_{m-1})+g(c_m),\ m>0}$$

It is denoted as follows:

$$\boxed{\Sigma\, g\Big|_{[a,c_k]} = \Sigma g\,(x_k)}$$

We also have:

$$h(x_m) = g(c_1) + g(c_2) + ... + g(c_m)\,.$$

In other words, this is an abbreviation of the *sigma notation* (which is also an abbreviation):

$$g(c_1) + g(c_2) + ... + g(c_m) = \sum_{i=1}^{m} g(c_i)\,.$$

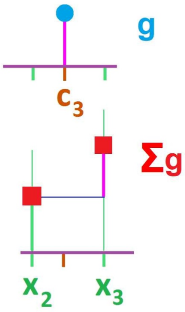
Here the original function g is evaluated over the interval, $[a, c_k]$, while the new function $h = \Sigma g$ is defined at x_k:

$$\Sigma\, g\Big|_{[a,c_k]} = \Sigma g\,(x_k)\,.$$

We can also use parentheses to separate functions from their inputs when it's not cumbersome:

$$\Sigma \left(g\Big|_{[a,c_k]} \right) = (\Sigma g)(x_k)\,.$$

The sum is computed recursively one segment at a time:



In this context, the secondary nodes are also known as the *sample points*.

In the examples, we chose the secondary nodes in a consistent way. First we choose to have equal increments:

$$h = \Delta x = \frac{b-a}{n}\,.$$

Furthermore, there are three main "schemes" for choosing secondary nodes. One is seen above: The secondary nodes are placed at the end of each interval. It is called the *right-end scheme*:

- Primary nodes: $x = a, a + h, a + 2h, ...$
- Secondary nodes: $c = a + h, a + 2h, ...$

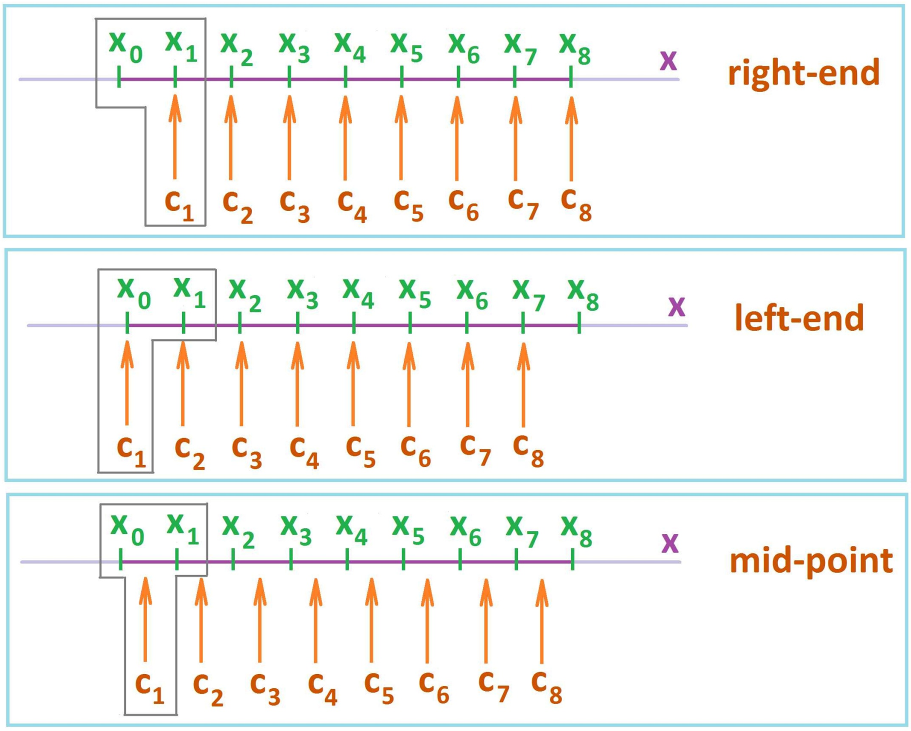
This is the *left-end scheme*:

- Primary nodes: $x = a, a + h, a + 2h, ...$
- Secondary nodes: $c = a, a + h, ...$

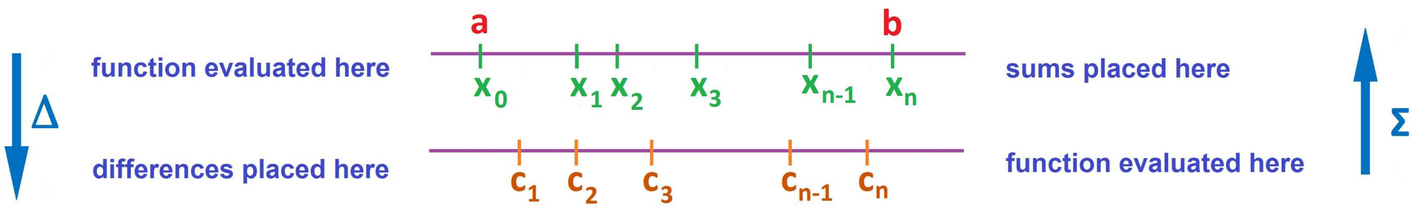
Another convenient choice is the *mid-point scheme*:

- Primary nodes: $x = a, a + h, a + 2h, ...$
- Secondary nodes: $c = a + h/2, a + 3h/2, ...$

They are illustrated below:



Below is the summary of our back-and-forth construction:



Now, even if the person didn't spend any time driving, the displacement still makes sense. It's zero!

We extend the meaning of the difference to include zero increments:

Definition 1.2.22: difference over zero-length interval

When the increment of x is zero:

$$\Delta x_k = x_k - x_{k-1} = 0,$$

the *difference* of a function g over this segment is defined to be zero:

$$\boxed{\Delta \left(g\big|_{[x_k,x_k]} \right) = 0}$$

We extend the meaning of the sum to include intervals of zero length.

Definition 1.2.23: sum over zero-length interval

The *sum* of a function g over an interval $[a, a]$ is defined to be zero:

$$\boxed{\Sigma \left(g\big|_{[a,a]} \right) = 0}$$

To capitalize on this idea even further, we extend the meaning of the difference and the sum to all *oriented* segments:

- a b The interval $[a, b]$ is positively oriented when $a < b$.
- b a The interval $[a, b]$ is negatively oriented when $a > b$.

In the latter case, the length of the interval is the distance from a to b, i.e., $b - a$, which is negative!

Definition 1.2.24: difference over negatively oriented interval

The *difference of a function f over a negatively oriented interval* $[x_k, x_{k-1}]$, $x_{k-1} < x_k$, *of a partition* is defined to be the negative of the difference over $[x_{k-1}, x_k]$:

$$\boxed{\Delta\left(f\big|_{[x_k, x_{k-1}]}\right) = -\Delta\left(f\big|_{[x_{k-1}, x_k]}\right)}$$

Definition 1.2.25: sum over negatively oriented interval

The *sum of a function g over a negatively oriented interval* $[b, a]$, $a < b$, over an augmented partition P of $[a, b]$, is defined to be the negative of the sum over P:

$$\boxed{\Sigma\left(g\big|_{[b,a]}\right) = -\Sigma\left(g\big|_{[a,b]}\right)}$$

Warning!

In anticipation of higher dimensional calculus, an oriented segment is *not* a vector.

We can also think of flipping an interval as multiplication by -1. If I stands for such an interval, we have:

$$\sum_{-I} f = -\sum_{I} f.$$

We next consider the simplest case.

Theorem 1.2.26: Sum of Constant Function

Suppose g is constant at the secondary nodes of a partition of $[a, b]$, i.e., $g(c_i) = p$ for all $i = 1, 2, ..., n$ and some real number p. Then

$$\boxed{\Sigma g\big|_{[a,b]} = p \cdot n}$$

Things are paired up:

Differential calculus	**Integral calculus**
f defined on the primary nodes	g defined on the secondary nodes
1. difference, Δf defined on the secondary nodes $\Delta f\,(c_k) = f(x_k) - f(x_{k-1})$	1. sum, Σg defined on the primary nodes $\Sigma g\,(x_k) = g(c_1) + ... + g(c_k)$
2.	2.

We continue with this table below.

1.3. The total value of a function: the Riemann sum

Let's generalize the examples from the last section.

Example 1.3.1: sampling location

First, we consider the trip with a broken speedometer.

We have a time interval $[a,b]$. In order to estimate our speed, we decide to look at the odometer several times during the trip, including the very beginning and the very end of it. Otherwise, the moments of time may be arbitrary:

$$a = x_0,\ x_1,\ x_2,\ \ldots,\ x_{n-1},\ x_n = b\,.$$

In other words, we have a partition of $[a,b]$, and the plan is to *sample the location*. The location is a function p defined on the whole interval, but now only its values at the primary nodes of the partition are recorded. We compute the difference quotient of this sequence producing the average velocity. Even though the velocity is a function v defined on the whole interval, we have now only its approximations assigned to the secondary nodes of the partition.

Example 1.3.2: sampling velocity

On the flip side, we consider the trip with a broken odometer.

We still have a time interval $[a,b]$. We split $[a,b]$ into smaller time intervals in an arbitrary manner. In other words, we have a partition of $[a,b]$, and the plan is to *sample the velocity*. During each of them we looked at the speedometer with the exact moments:

$$c_1,\ c_2,\ c_3,\ \ldots,\ c_{n-1},\ c_n\,.$$

They are just a matter of bookkeeping. The velocity is a function v defined on the whole interval, but now only its values at the secondary nodes of the partition are recorded. We multiply the velocity by the length of the current time interval producing the approximate displacement. Then we add them together (the "Riemann sum"). Even though the displacement is a function p defined on the whole interval, we have now only its approximations assigned to the primary nodes of the partition.

This is an example of such a computation:

time intervals (hours):	$[0,2]$	$[2,4]$	$[4,6]$	$[6,8]$
time (hours):	$c_1 = 1$	$c_2 = 4$	$c_3 = 5$	$c_4 = 6$
velocity (miles/hour):	60	100	-80	-80
displacement (miles):	$60 \cdot 2$	$100 \cdot 2$	$-80 \cdot 2$	$-80 \cdot 2$
	$= 120$	$= 200$	$= -160$	$= -160$
total displacement (miles):	120	$120 + 200$	$340 - 160$	$180 - 160$
	$= 120$	$= 340$	$= 180$	$= 20$

We have the displacement as a function p of time but we can also assign these numbers to the primary nodes:

time (hours):	$x_0 = 0$	$x_1 = 2$	$x_2 = 4$	$x_3 = 6$	$x_4 = 8$
total displacement (miles):	0	120	340	180	20

Of course, if we need the location, we need our initial location first, i.e., the value of $p(a)$, in order to start the computation; we assumed it to be 0 above. If it's 10, we have:

time (hours):	$x_0 = 0$	$x_1 = 2$	$x_2 = 4$	$x_3 = 6$	$x_4 = 8$
total displacement (miles):	10	130	350	190	30

Exercise 1.3.3

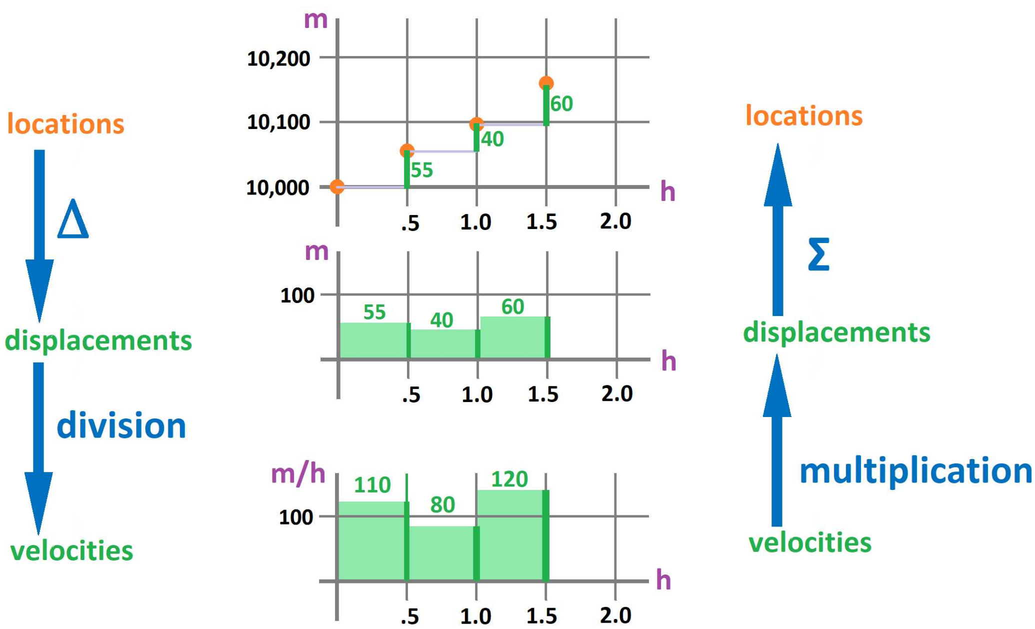
What is the approximate displacement if the initial position was -50? What is the approximate displacement if the position at time 2 was 20?

Exercise 1.3.4

Compute the approximate displacement for the same sampling data as in the example but for a different (unequal) partition:

time intervals (hours):	$[0,1]$	$[1,4]$	$[4,5]$	$[5,8]$

So, in comparison to the last section, the new step is the division or multiplication by the time increment Δx:



Let's provide precise definitions. Just as in the last section, there will be no restrictions whatsoever on the nodes of the partition.

Definition 1.3.5: difference quotient of function

Suppose a function $y = f(x)$ is defined at the primary nodes $x_k,\ k = 0, 1, 2, ..., n$, of a partition. Then the *difference quotient* of f is a function defined at the

secondary nodes $c_k,\ k=1,2,...,n$ of the partition as the following fraction:

$$\frac{\Delta f}{\Delta x}(c_k)=\frac{f(x_{k+1})-f(x_k)}{x_{k+1}-x_k}=\frac{f(x_k+\Delta x_k)-f(x_k)}{\Delta x_k}$$

It is the difference of f divided by Δx_k.

It is the relative change – the *rate* of change – of the two sequences.

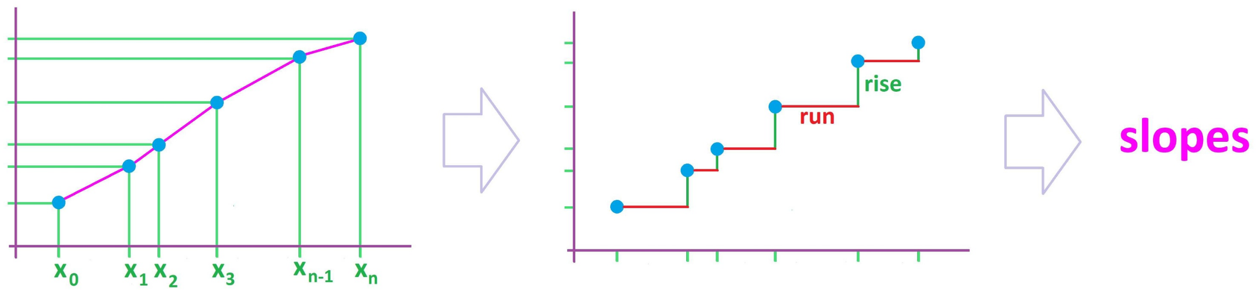
Let's remember that here the original function f is evaluated over the interval, $[x_{k-1},x_k]$, while the new function Δf is defined at c_k:

$$\frac{\Delta}{\Delta x}\, f\Big|_{[x_{k-1},x_k]}=\frac{\Delta f}{\Delta x}\,(c_k)\,.$$

We can also use parentheses to separate functions from their inputs when it's not cumbersome:

$$\frac{\Delta}{\Delta x}\left(f\Big|_{[x_{k-1},x_k]}\right)=\left(\frac{\Delta f}{\Delta x}\right)(c_k)\,.$$

Where in the last section we looked for the rises, we find the slopes now. On the graph, we can see how each consecutive pair of points produces a line segment and the difference quotient is the *slope* of the line connecting them:

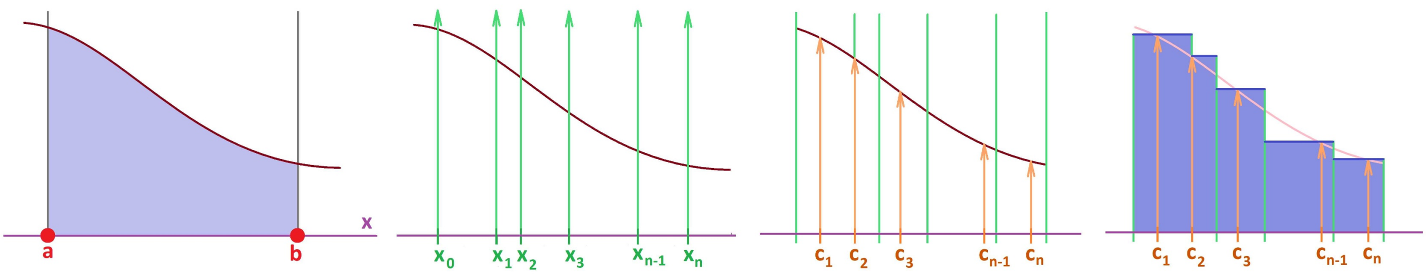


Now in reverse.

The example of displacement is an especially important case of the sum of a function computed from another one as follows:

$$g(c_k)=f(c_k)\Delta x_k\,.$$

The construction is the same as in the last section, but the increments of x make their appearance:



These are our two interpretations of the same construction:

	motion	geometry
Δx_k	the length of the kth time interval	the width of the kth rectangle
$f(c_k)$	the velocity during the kth time interval	the height of the kth rectangle
$g(c_k)=f(c_k)\Delta x_k$	the displacement during the kth time interval	the area of the kth rectangle

Warning!

Even though we often say that each such term is the "area" of the rectangle, this quantity is truly the area only if the unit of both the x- and the y-axis is that of length.

We define a special kind of sum.

Definition 1.3.6: Riemann sum

Suppose a function f defined at the secondary nodes of an augmented partition P of an interval $[a,b]$. The *Riemann sum* of f is the function h defined on the primary nodes as the sum of the sequence:

$$f(c_1)\Delta x_1,\ f(c_2)\Delta x_2,\ \dots,\ f(c_n)\Delta x_n\,.$$

In other words, it is the following sequence defined recursively:

$$h(x_0)=0,\quad h(x_k)=h(x_{k-1})+f(c_k)\Delta x_k,\ k=1,2,\dots,n\,.$$

It is denoted as follows:

$$\boxed{\Sigma\, f\cdot\Delta x\Big|_{[a,c_k]}=\Sigma f\cdot\Delta x\,(x_k)=f(c_1)\Delta x_1+f(c_2)\Delta x_2+\dots+f(c_k)\Delta x_k}$$

In sigma notation, it is:

$$\sum_{i=1}^{k} f(c_i)\Delta x_i\,.$$

Here the original function f is evaluated over the interval, $[a,c_k]$, while the new function $h=\Sigma f\cdot\Delta x$ is defined at x_k:

$$\Sigma\, f\cdot\Delta x\Big|_{[a,c_k]}=\Sigma f\cdot\Delta x\,(x_k)\,.$$

We can also use parentheses to separate functions from their inputs when it's not cumbersome:

$$\Sigma\left(f\cdot\Delta x\Big|_{[a,c_k]}\right)=(\Sigma f\cdot\Delta x)(x_k)\,.$$

At first we will concentrate on the Riemann sum over the *whole interval*.

Let's consider some specific examples of Riemann sums. We assume again that the lengths of the intervals are equal:

$$\Delta x_i=\Delta x=(b-a)/n\,.$$

Example 1.3.7: x^2

Let $f(x)=x^2$ over the interval $[0,1]$ with $n=4$.

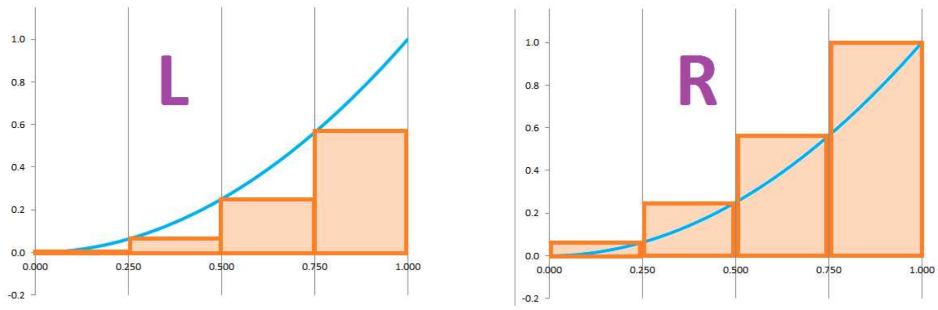
Then $\Delta x=1/4$ and the interval is subdivided as before and we have five values of the function to

work with:

	$\bullet$	$--$	$\vert$	$--$	$\vert$	$--$	$\vert$	$--$	$\bullet$	
partition: $x=$	0		$1/4$		$1/2$		$3/4$		1	
values: $x^2=$	0		$1/16$		$1/4$		$9/16$		1	
left-end RS $=$	$0\cdot 1/4$	$+$	$1/16\cdot 1/4$	$+$	$1/4\cdot 1/4$	$+$	$9/16\cdot 1/4$			≈ 0.22
right-end RS $=$			$1/16\cdot 1/4$	$+$	$1/4\cdot 1/4$	$+$	$9/16\cdot 1/4$	$+$	$1\cdot 1/4$	≈ 0.47

The simplest choices for the secondary nodes are the *left-end* or the *right-end* points of the intervals.

This fully algebraic construction can be visualized:



Exercise 1.3.8

What is your best estimate of the curved area based on the data provided?

Exercise 1.3.9

Repeat the computations for (a) n=8, (b) $[a,b]=[-1,0]$, (c) $f(x)=x^3$.

This is the left-end partition:

$$a=x_0 \blacksquare x_1 \blacksquare x_2 \quad \ldots \quad x_{n-1} \blacksquare x_n=b \qquad \text{primary nodes}$$

$$a=c_1 \blacksquare c_2 \blacksquare c_3 \quad \ldots \quad c_n \blacksquare \circ \qquad \text{secondary nodes}$$

The table shows the nth left-end Riemann sum and the nth right-end Riemann sum of a function f on $[a,b]$:

i	0		1		$\ldots$		$n-1$		n
$[x_i,x_{i+1}]$	$\vert$	$--$	$\vert$	$-$	$\ldots$	$-$	$\vert$	$--$	$\vert$
x_i	a		$a+\Delta x$		$\ldots$		$a+(n-1)\,\Delta x$		b
$f(x)$	$f(x_0)$		$f(x_1)$		$\ldots$		$f(x_{n-1})$		$f(x_n)$
left-end RS $=$	$f(x_0)\cdot\Delta x$	$+$	$f(x_1)\cdot\Delta x$	$+$	$\ldots$	$+$	$f(x_{n-1})\cdot\Delta x$		
right-end RS $=$			$f(x_1)\cdot\Delta x$	$+$	$\ldots$	$+$	$f(x_{n-1})\cdot\Delta x$	$+$	$f(x_n)\cdot\Delta x$

The formula for such a sequence of secondary nodes is the same for both:

$$c_i=a+i\,\Delta x\,,$$

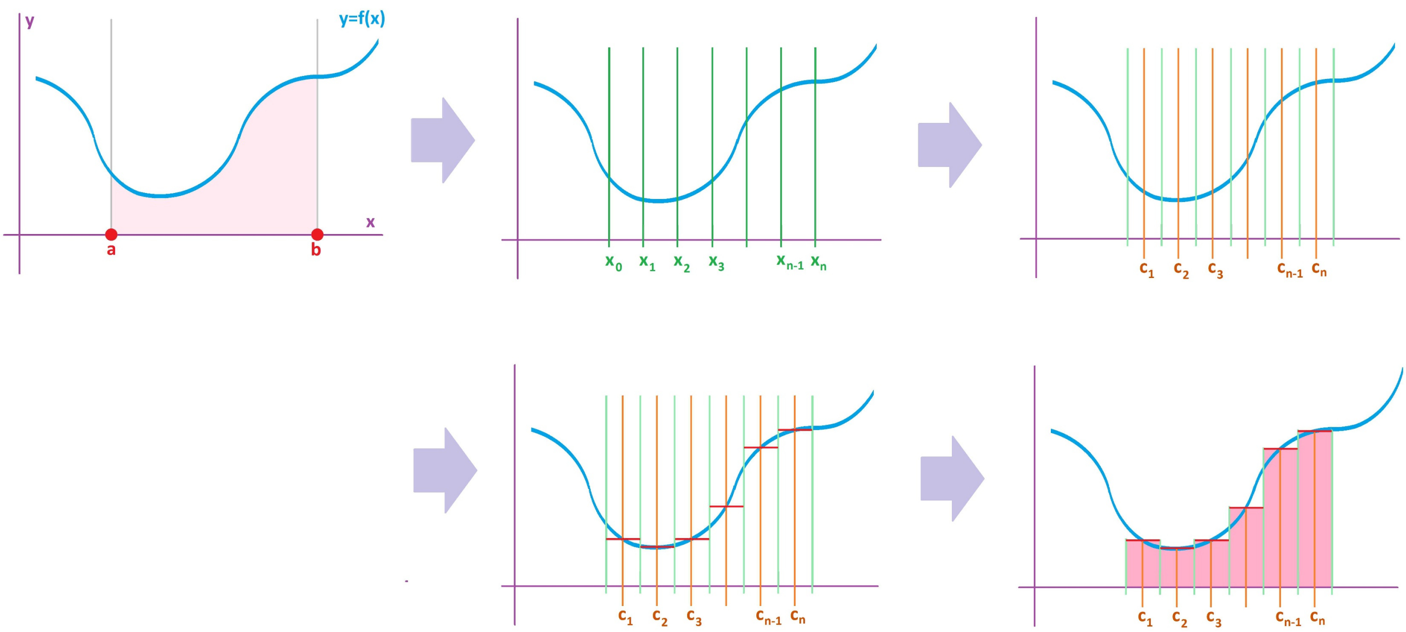
but the indices run over different sets:

- $i = 0, 1, 2, ..., n-1$ for the left-end points;
- $i = 1, 2, ..., n$ for the right-end points.

The secondary nodes can also be chosen to be the *mid-points* of the intervals:

i	0		1		2	...	$n-1$		n
$[x_i, x_{i+1}]$	$\vert$	$-\cdot-$	$\vert$	$-\cdot-$	$\vert$	$-$ ... $-$	$\vert$	$-\cdot-$	$\vert$
c_i		$a+\frac{1}{2}\Delta x$		$a+\frac{3}{2}\Delta x$		...		$a+\frac{2n-1}{2}\Delta x$	
$f(x)$		$f(c_0)$		$f(c_1)$		...		$f(c_n)$	
mid-point RS $=$		$f(c_1)\cdot\Delta x$	$+$	$f(c_2)\cdot\Delta x$	$+$	...	$+$	$f(c_n)\cdot\Delta x$	

The illustration shows the nth mid-point Riemann sum of f on $[a, b]$:



The single formula representation of the mid-points is

$$c_i = a + \left(i + \tfrac{1}{2}\right)\Delta x,\ \ i = 0, 1, 2, ..., n-1\,.$$

Example 1.3.10: negative area

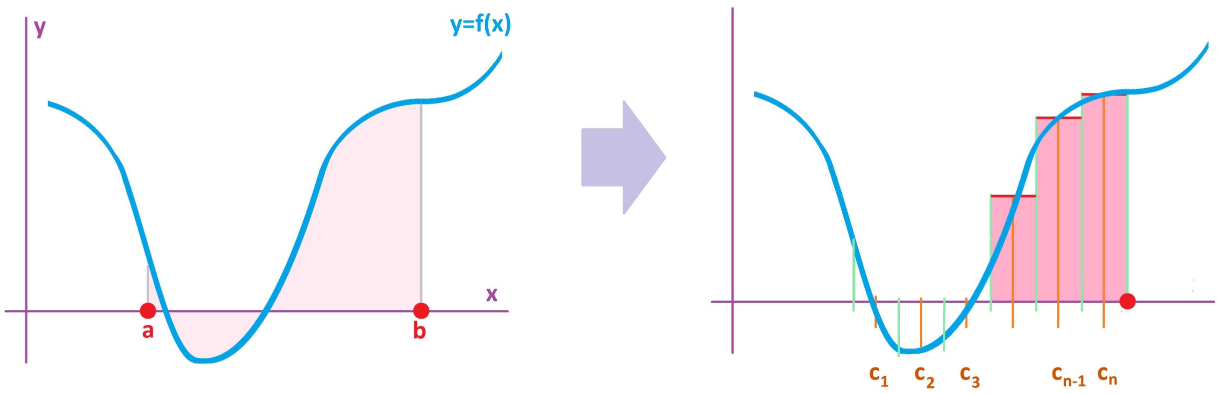
When the region lies within the Euclidean plane, the ith term in all three Riemann sums is:

$$\text{the area of } i\text{th rectangle} = \underbrace{f(c_i)}_{\text{height of rectangle}} \times \overbrace{\Delta x_i}^{\text{width of rectangle}}$$

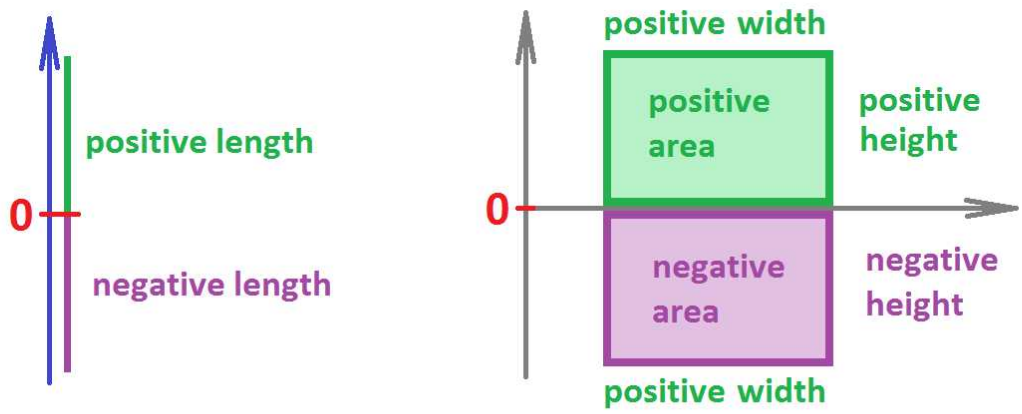
The Riemann sum is the total area of the rectangles. But what if the values of f are *negative*,

$$f(c_i) < 0?$$

Nothing in the definition prevents that:



What is the meaning of each term $f(c_i)\,\Delta x$ then? Just imagine again that f is the velocity. Then $f(c_i)\,\Delta x$ is still distance "covered" but with the *negative* speed, which means that you are moving in the opposite direction! Then your *displacement* is negative. As for the area metaphor, since $\Delta x > 0$, the meaning is determined by the meaning of $f(c_i)$:



We follow up to the last section and extend the meaning of the Riemann sum to include intervals of zero length.

Definition 1.3.11: Riemann sum over zero-length interval

The *Riemann sum* of a function f over an interval $[a, a]$ is defined to be zero:

$$\boxed{\Sigma\left(f \cdot \Delta x\Big|_{[a,a]}\right) = 0}$$

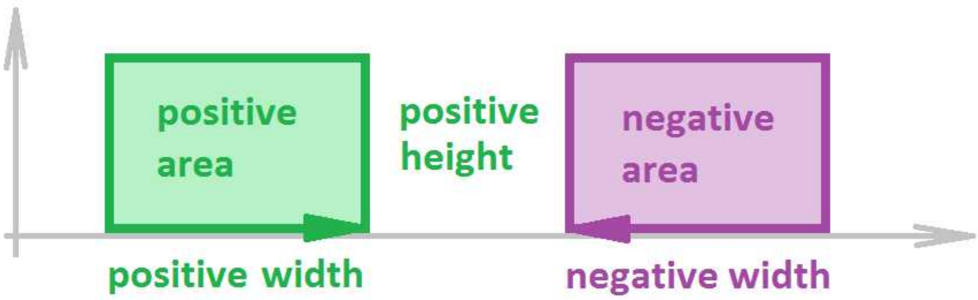
Definition 1.3.12: Riemann sum over negatively oriented interval

The *Riemann sum of a function f over a negatively oriented interval* $[b, a]$, $a \le b$, over an augmented partition P of $[a, b]$, is defined to be the negative of the Riemann sum over P:

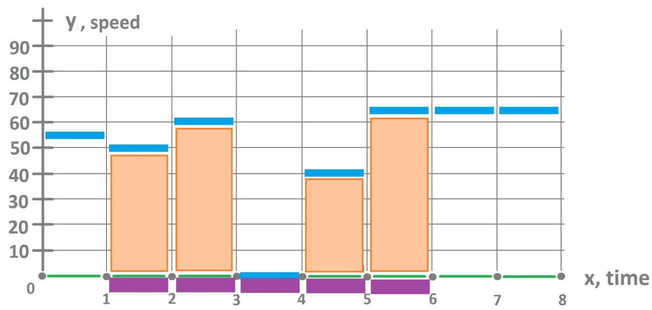
$$\boxed{\Sigma\left(f \cdot \Delta x\Big|_{[b,a]}\right) = -\Sigma\left(f \cdot \Delta x\Big|_{[a,b]}\right)}$$

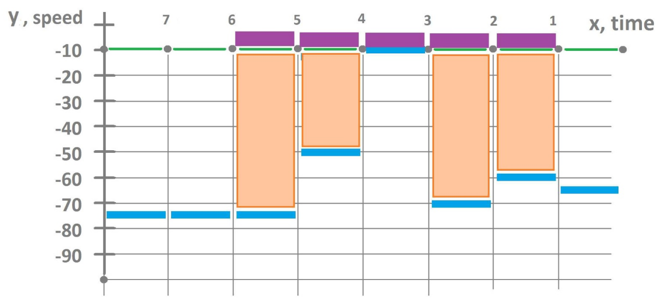
Example 1.3.13: negative width

When the region lies within the Euclidean plane, the width of the corresponding rectangle in the Riemann sum is also negative, and so is its area when the function is positive:



The notion of a *negative area* is justified by the fact that the rectangle is negatively oriented... The motion metaphor might explain it better though. It is as if *time is reversed* (or a videotape goes backward), so that the direction of motion is opposite, and all gains are reversed:





We next consider the simplest case of a Riemann sum.

Theorem 1.3.14: Riemann Sum of Constant Function

Suppose f is constant at the secondary nodes of a partition of $[a,b]$, i.e., $f(c_i)=p$ for all $i=1,2,...,n$ and some real number p. Then

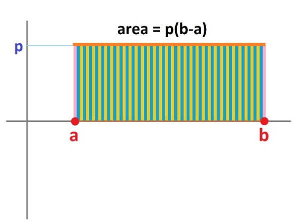
$$\boxed{\Sigma\left(f\cdot\Delta x\Big|_{[a,b]}\right)=p(b-a)}$$

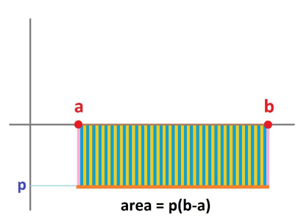
Proof.

Since $f(c_i)=p$ for all i, the Riemann sum is equal to:

$$\begin{aligned}\Sigma f\cdot\Delta x &= f(c_1)\,\Delta x_1+f(c_2)\,\Delta x_2+...+f(c_n)\,\Delta x_n\\ &= p\,\Delta x_1+p\,\Delta x_2+...+p\,\Delta x_n\\ &= p\big[\Delta x_1+\Delta x_2+...+\Delta x_n\big]\\ &= p\big[(x_1-x_0)+(x_2-x_1)+(x_3-x_2)+...+(x_n-x_{n-1})\big]\\ &= p(b-a)\,.\end{aligned}$$

Once we zoomed out, we recognize that the Riemann sum represents the rectangle with width $b-a$ and height p:

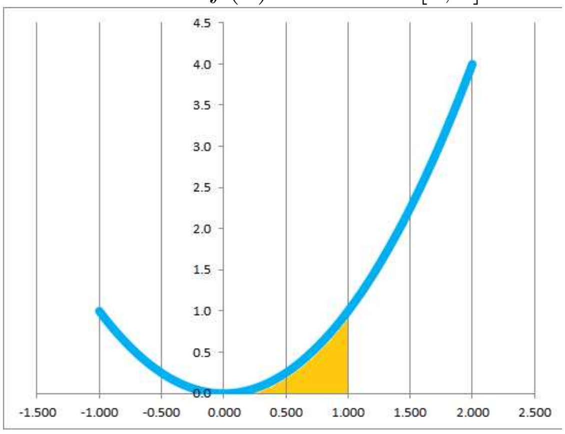




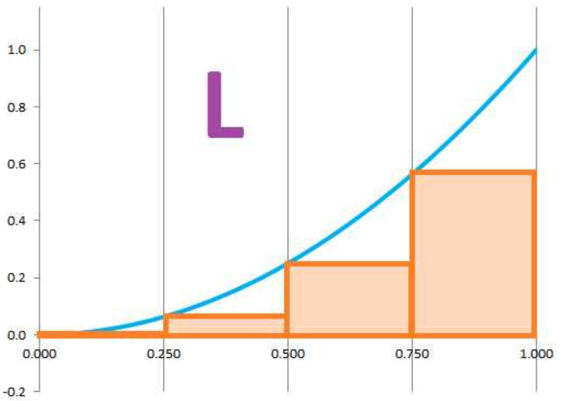
Except, it is negative when p is negative.

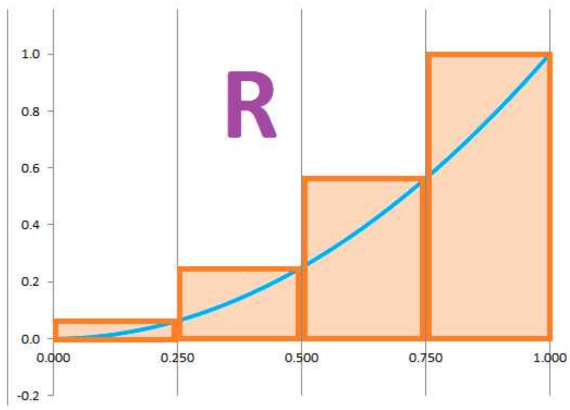
Example 1.3.15: Riemann sums are functions

Let's consider *all* of the Riemann sums of of $f(x) = x^2$ over $[0, 1]$:



We choose the number of intervals to be $n = 4$ with equal intervals of length $h = 1/4$, and we choose, as the secondary nodes, the left-end or the right-end of each interval:



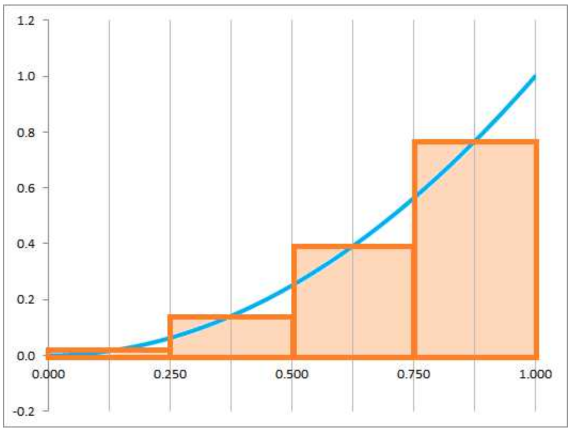


At those points, the function is evaluated. This is the computation of the left-end Riemann sum L_4:

	$\bullet$	$--$	$\vert$	$--$	$\vert$	$--$	$\vert$	$--$	$\bullet$	
x	0		$1/4$		$1/2$		$3/4$		1	
x^2	0		$1/16$		$1/4$		$9/16$		1	
L_4	$0 \cdot 1/4$	$+$	$1/16 \cdot 1/4$	$+$	$1/4 \cdot 1/4$	$+$	$9/16 \cdot 1/4$			≈ 0.22
$\Sigma_{[0,0]}$	$0 \cdot 1/4$									$= 0$
$\Sigma_{[0,1/4]}$	$0 \cdot 1/4$	$+$	$1/16 \cdot 1/4$							≈ 0.04
$\Sigma_{[0,1/2]}$	$0 \cdot 1/4$	$+$	$1/16 \cdot 1/4$	$+$	$1/4 \cdot 1/4$					≈ 0.10
$\Sigma_{[0,3/4]}$	$0 \cdot 1/4$	$+$	$1/16 \cdot 1/4$	$+$	$1/4 \cdot 1/4$	$+$	$9/16 \cdot 1/4$			≈ 0.22

We, furthermore, realize that we are computing the Riemann sum *function* for this augmented partition. Its four values are shown in bottom of the table.

We can also choose the mid-points as the secondary nodes:



This is the computation of the mid-point Riemann sum M_4 for the same integral:

	$\bullet$ $--$	$\vert$	$--$	$\vert$	$--$	$\vert$	$--$	$\bullet$
x	$1/8$		$3/8$		$5/8$		$7/8$	
$f(x)=x^2$	$(1/8)^2$		$(3/8)^2$		$(5/8)^2$		$(7/8)^2$	
M_4	$(1/8)^2 \cdot 1/4$	$+$	$(3/8)^2 \cdot 1/4$	$+$	$(5/8)^2 \cdot 1/4$	$+$	$(7/8)^2 \cdot 1/4$	$=0.328125$
$\Sigma_{[0,1/8]}$	$(1/8)^2 \cdot 1/4$							≈ 0.004
$\Sigma_{[0,3/8]}$	$(1/8)^2 \cdot 1/4$	$+$	$(3/8)^2 \cdot 1/4$					≈ 0.040
$\Sigma_{[0,5/8]}$	$(1/8)^2 \cdot 1/4$	$+$	$(3/8)^2 \cdot 1/4$	$+$	$(5/8)^2 \cdot 1/4$			≈ 0.230
$\Sigma_{[0,7/8]}$	$(1/8)^2 \cdot 1/4$	$+$	$(3/8)^2 \cdot 1/4$	$+$	$(5/8)^2 \cdot 1/4$	$+$	$(7/8)^2 \cdot 1/4$	≈ 0.328

It is much closer than L_4 to the true value of the integral, $1/3$.

Here is the summary of the development up to this point:

Differential calculus	**Integral calculus**
f defined on the primary nodes	g defined on the secondary nodes
1. difference, Δf defined on the secondary nodes $\Delta f\,(c_k)=f(x_k)-f(x_{k-1})$	1. sum, Σg defined on the primary nodes $\Sigma g\,(x_k)=g(c_1)+...+g(c_k)$
divided by Δx	Δx is factored in each term
2. difference quotient, $\dfrac{\Delta f}{\Delta x}$ defined on the secondary nodes $\dfrac{\Delta f}{\Delta x}(c_k)=\dfrac{f(x_k)-f(x_{k-1})}{\Delta x}$	2. Riemann sum, $\Sigma f \cdot \Delta x$ defined on the primary nodes $\Sigma f \cdot \Delta x\,(x_k)=f(c_1)\Delta x+...+f(c_k)\Delta x$
3.	3.

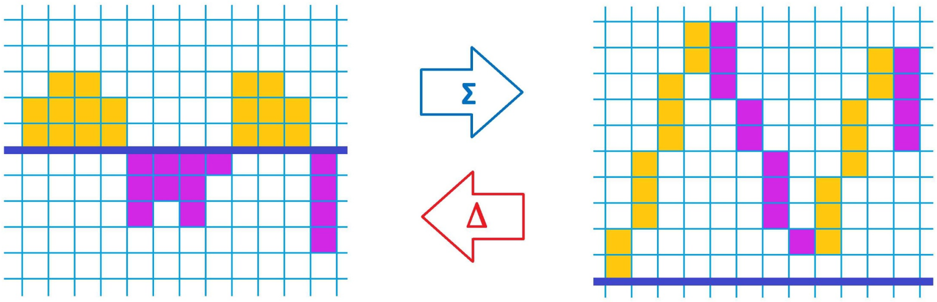
The two columns are constructed independently but can easily be linked together.

1.4. The Fundamental Theorem of Calculus

As we know, addition and subtraction cancel each other.

Repeated addition and subtraction cancel each other, too.

In other words, the sum and the difference – as operations on sequences – cancel each other. We illustrate this idea with the following:



As you can see, the sum stacks up the terms of the sequence on top of each other while the difference takes this back down.

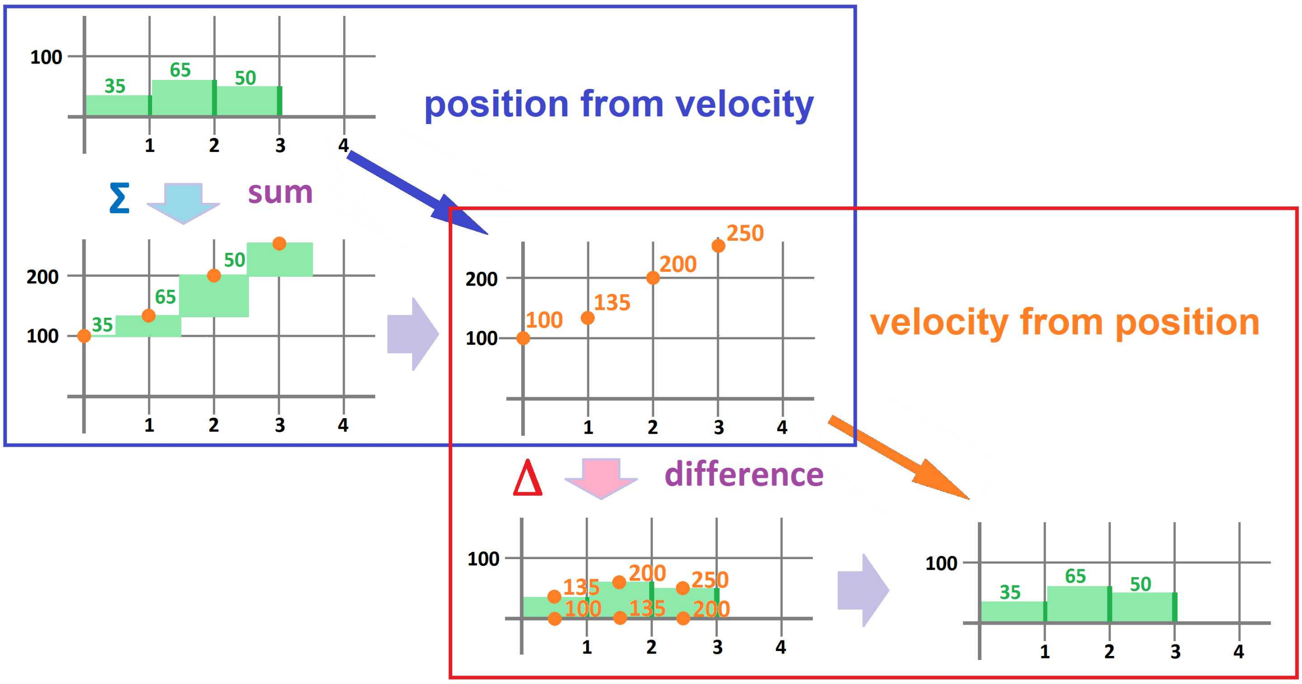
What do the sum and the difference have to do with the Riemann sums and the difference quotients? Just plug in $\Delta x_k = 1$ and you get the former from the latter, respectively. The relation between the Riemann sums and the difference quotients is *fundamental*.

Example 1.4.1: cancellation

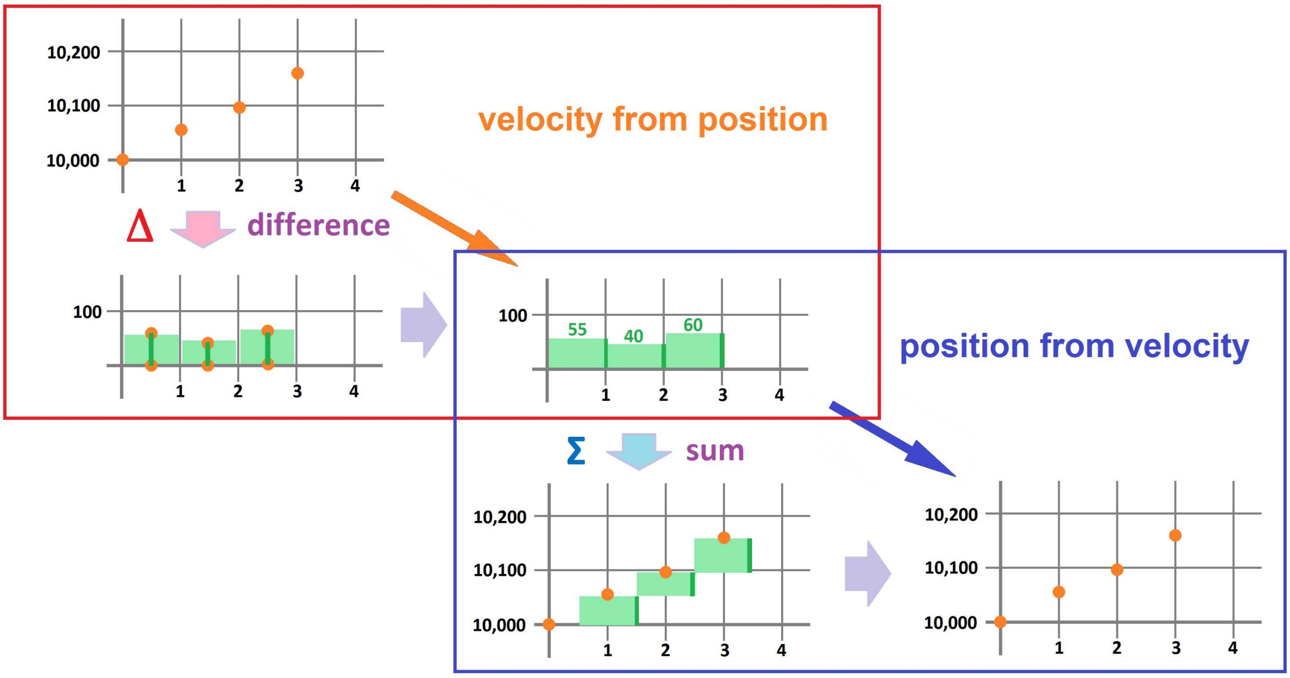
We know how to get the velocity from the location – and the location from the velocity. Of course, executing these two operations consecutively should bring us back where we started.

We now take another look at the *two* computations about motion – a broken odometer and a broken speedometer – presented earlier.

First, below we see how the velocities are used to acquire the displacements via the sums, but we also discover that we can get the former from the latter via the differences:



Second, below we see how the positions are used to acquire the velocities via the differences, but we also discover that we can get the former from the latter via the sums:



We knew this would happen: After all, the position and the velocity aren't really derived from each other but co-exists as two attributes of the same motion.

Suppose once again that we have an augmented partition P of an interval $[a, b]$ with n intervals and

- the primary nodes: $x_k,\ k = 0, 1, ..., n$
- the increments: $\Delta x_k = x_k - x_{k-1},\ i = 1, 2, ..., n$
- the secondary nodes: $c_k,\ i = 1, 2, ..., n$ in $[x_{k-1}, x_k]$

There are two parts:

1. **Sum:** Suppose we have a function g defined at the secondary nodes c_k. We compute its sum with a variable right end, x_k. We assign these values to these primary nodes. This defines a new function,

G, defined on the primary nodes. The new function can be written explicitly:

$$G(x_k) = \Sigma g\,(x_k) = g(c_1) + g(c_2) + ... + g(c_k)\,,$$

or computed recursively:

$$G(x_k) = G(x_{k-1}) + g(c_k)\,.$$

2. **Difference:** Suppose we have a function F defined at the primary nodes of the partition. Just as always, its difference is computed over *each* interval $[x_{k-1}, x_k]$ of the partition, as follows:

$$F(x_k) - F(x_{k-1}),\ \ k = 1, 2..., n\,.$$

We then have a number – representing the rise – for each k. This defines a new function, f, on the secondary nodes. The new function is computed by an explicit formula:

$$f(c_k) = F(x_k) - F(x_{k-1})\,.$$

The first question we would like to answer is:

▶ *What is the difference of the sum?*

We have g defined at the secondary nodes of the partition and its sum G defined recursively at the primary nodes:

$$G(x_k) = G(x_{k-1}) + g(c_k)\,.$$

Also, the difference f of G is defined at the secondary nodes by:

$$f(c_k) = G(x_k) - G(x_{k-1})\,.$$

We *substitute* the latter into the former:

$$f(c_k) = G(x_k) - G(x_{k-1}) = g(c_k)\,.$$

So, the answer is, *the original function.*

The result takes the following compact form:

Theorem 1.4.2: Fundamental Theorem of Discrete Calculus I

1. *The difference of the sum of g is g:*

$$\boxed{\Delta\,(\Sigma g) = g}$$

2. *The difference quotient of the Riemann sum of g is g:*

$$\boxed{\frac{\Delta\,(\Sigma g \cdot \Delta x)}{\Delta x} = g}$$

The two operations *cancel* each other!

Proof.

For the second part, we need a slight modification of the above argument. We have g defined at the secondary nodes of the partition and its (variable-end) Riemann sum G defined recursively at the primary nodes:

$$G(x_k) = G(x_{k-1}) + g(c_k)\,\Delta x_k\,.$$

Also, the difference quotient f of G is defined at the secondary nodes by:

$$f(c_k) = \frac{G(x_k) - G(x_{k-1})}{\Delta x_k}\,.$$

We substitute the latter into the former:

$$\begin{aligned} f(c_k) &= \frac{G(x_k) - G(x_{k-1})}{\Delta x_k} \\ &= \frac{g(c_k)\,\Delta x_k}{\Delta x_k} \\ &= g(c_k)\,. \end{aligned}$$

The second question we would like to answer is:

- *What is the sum of the difference?*

We have F defined at the primary nodes of the partition and its difference f defined at the secondary nodes:

$$f(c_k) = F(x_k) - F(x_{k-1})\,.$$

Also, the sum G of f is defined at the primary nodes by:

$$G(x_k) = G(x_{k-1}) + f(c_k)\,.$$

We *substitute* the former into the latter:

$$G(x_k) - G(x_{k-1}) = f(c_k) = F(x_k) - F(x_{k-1})\,.$$

Furthermore,

$$\begin{aligned} G(x_k) - G(x_0) &= \big[G(x_k) - G(x_{k-1})\big] + \big[G(x_{k-1}) - G(x_{k-2})\big] + ... + \big[G(x_1) - G(x_0)\big] \\ &= \big[F(x_k) - F(x_{k-1})\big] + \big[F(x_{k-1}) - F(x_{k-2})\big] + ... + \big[F(x_1) - F(x_0)\big] \\ &= F(x_k) - F(x_0)\,. \end{aligned}$$

So, the answer is, *the original function plus a constant.*

Theorem 1.4.3: Fundamental Theorem of Discrete Calculus II

1. *The sum of the difference of F is $F + C$, where C is a constant:*

$$\boxed{\Sigma\,(\Delta F) = F + C}$$

2. *The Riemann sum of the difference quotient of F is $F + C$, where C is a constant:*

$$\boxed{\Sigma\left(\frac{\Delta F}{\Delta x}\right)\Delta x = F + C}$$

Proof.

For the second part, we need a slight modification of the above argument. We have F defined at the

primary nodes of the partition and its difference quotient f defined at the secondary nodes:

$$f(c_k) = \frac{F(x_k) - F(x_{k-1})}{\Delta x_k}\,.$$

Also, the Riemann sum G of f is defined at the primary nodes by:

$$G(x_k) = G(x_{k-1}) + f(c_k)\,\Delta x_k\,.$$

We substitute the former into the latter:

$$\begin{aligned} G(x_k) - G(x_{k-1}) &= f(c_k)\,\Delta x_k \\ &= \frac{F(x_k) - F(x_{k-1})}{\Delta x_k}\,\Delta x_k \\ &= F(x_k) - F(x_{k-1})\,. \end{aligned}$$

Furthermore,

$$\begin{aligned} G(x_k) - G(x_0) &= \big[G(x_k) - G(x_{k-1})\big] + \big[G(x_{k-1}) - G(x_{k-2})\big] + ... + \big[G(x_1) - G(x_0)\big] \\ &= \big[F(x_k) - F(x_{k-1})\big] + \big[F(x_{k-1}) - F(x_{k-2})\big] + ... + \big[F(x_1) - F(x_0)\big] \\ &= F(x_k) - F(x_0)\,. \end{aligned}$$

The two operations – almost – cancel each other, again! They don't cancel fully because differentiation isn't a one-to-one function.

In summary, these are the operations involved:

$$\begin{array}{rcccc} \text{difference quotient, } \dfrac{\Delta f}{\Delta x}: & \to & \text{subtraction} & \to & \text{division} \\ & & & & \downarrow \\ \text{Riemann sums, } \Sigma g \cdot \Delta x: & \leftarrow & \text{addition} & \leftarrow & \text{multiplication} \end{array}$$

We carry out these four operations consecutively. The multiplication and division by Δx cancel each other first:

$$\begin{array}{rcc} \text{difference, } \Delta f: & \to & \text{subtraction} \\ & & \downarrow \\ \text{sum, } \Sigma g: & \leftarrow & \text{addition} \end{array}$$

That's another cancellation!

Example 1.4.4: Fundamental Theorem with spreadsheet

For complex data, we use a spreadsheet with the formulas presented in Chapters 2DC-2 and 2DC-3. From a function to its Riemann sum:

```
=R[-1]C+RC[-1]*R1C2
```

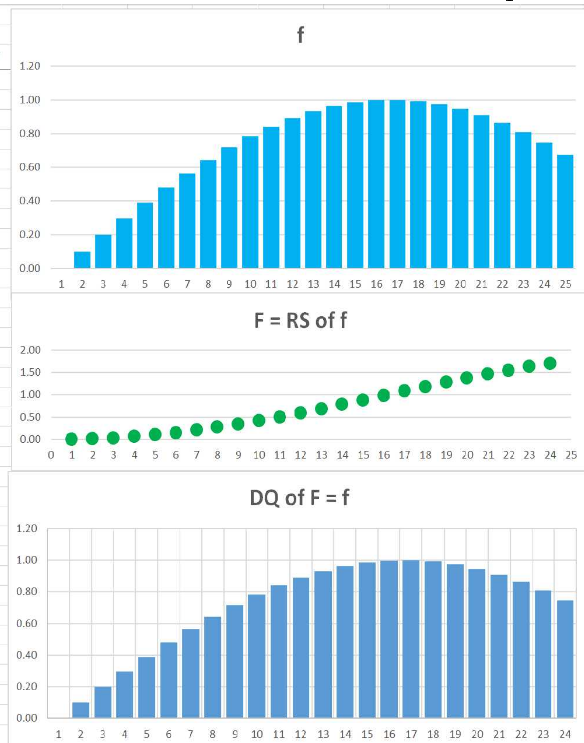
From a function to its difference quotient:

```
=(RC[-1]-R[-1]C[-1])/R1C2
```

What if we execute the two computations consecutively?

In this order first, from a function to its Riemann sum to the difference quotient of the latter:

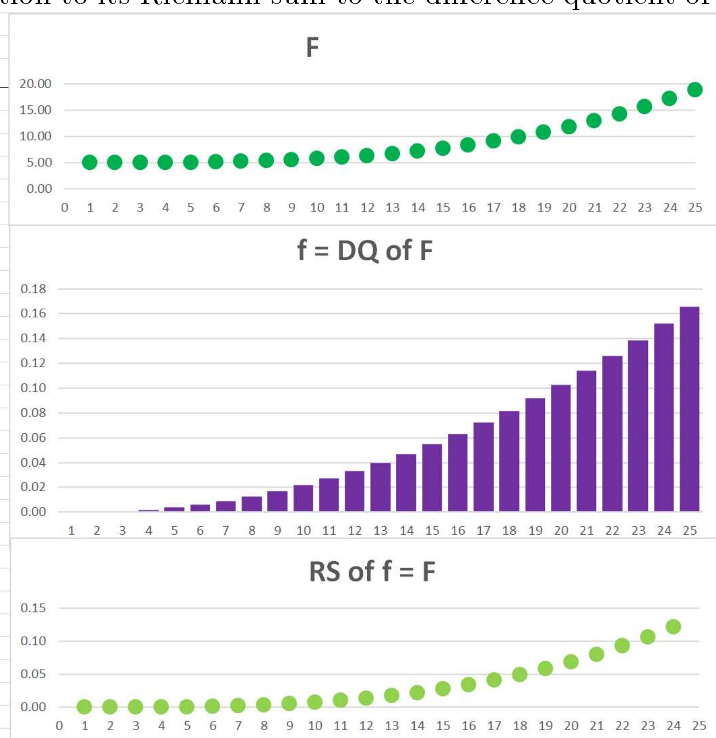
h			
0.10			
x	**f**	**F = RS of f**	**DQ of F = f**
0.00	0.00	0.00	
0.10	0.10	0.00	0.00
0.20	0.20	0.01	0.10
0.30	0.30	0.03	0.20
0.40	0.39	0.06	0.30
0.50	0.48	0.10	0.39
0.60	0.56	0.15	0.48
0.70	0.64	0.20	0.56
0.80	0.72	0.27	0.64
0.90	0.78	0.34	0.72
1.00	0.84	0.42	0.78
1.10	0.89	0.50	0.84
1.20	0.93	0.59	0.89
1.30	0.96	0.68	0.93
1.40	0.99	0.78	0.96
1.50	1.00	0.88	0.99
1.60	1.00	0.98	1.00
1.70	0.99	1.08	1.00
1.80	0.97	1.18	0.99
1.90	0.95	1.27	0.97
2.00	0.91	1.37	0.95
2.10	0.86	1.46	0.91
2.20	0.81	1.55	0.86
2.30	0.75	1.63	0.81
2.40	0.68	1.70	0.75



It's the same function!

Now in the opposite order, from a function to its Riemann sum to the difference quotient of the latter:

h			
0.10			
x	**F**	**f = DQ of F**	**RS of f = F**
0.00	5.00		
0.10	5.00	0.00	0.00
0.20	5.01	0.00	0.00
0.30	5.03	0.00	0.00
0.40	5.06	0.00	0.00
0.50	5.13	0.01	0.00
0.60	5.22	0.01	0.00
0.70	5.34	0.01	0.00
0.80	5.51	0.02	0.00
0.90	5.73	0.02	0.01
1.00	6.00	0.03	0.01
1.10	6.33	0.03	0.01
1.20	6.73	0.04	0.01
1.30	7.20	0.05	0.02
1.40	7.74	0.05	0.02
1.50	8.38	0.06	0.03
1.60	9.10	0.07	0.03
1.70	9.91	0.08	0.04
1.80	10.83	0.09	0.05
1.90	11.86	0.10	0.06
2.00	13.00	0.11	0.07
2.10	14.26	0.13	0.08
2.20	15.65	0.14	0.09
2.30	17.17	0.15	0.11
2.40	18.82	0.17	0.12

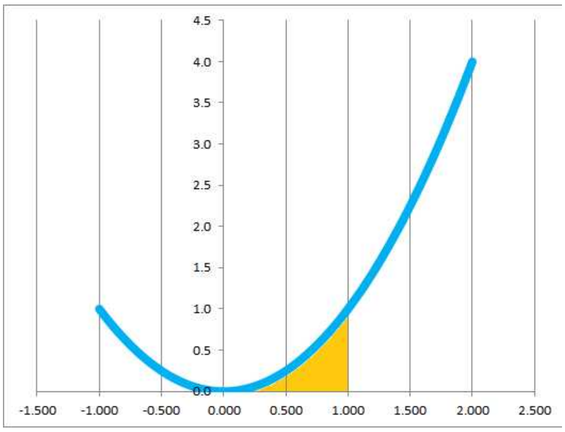


It's the same function shifted down!

1.5. How to approximate areas and displacements

Example 1.5.1: estimate areas

Let's estimate – in several ways – the area under the graph of $y = f(x) = x^2$ that lies above the interval $[0,1]$.

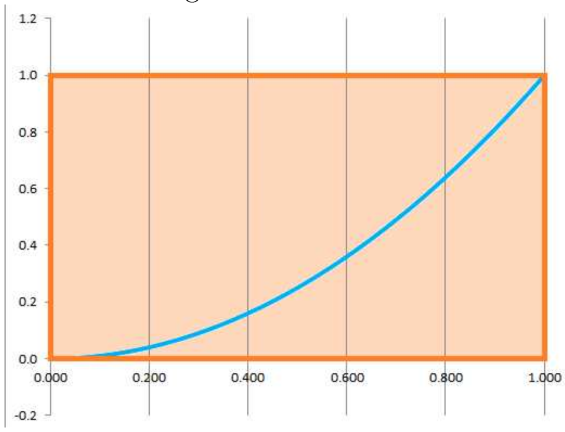


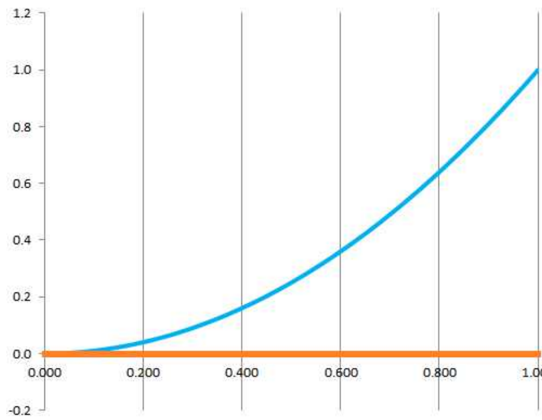
As before, we will *sample* our function at several values of x, a total of n of times.

Let's start with $n = 1$ and pick the *right end* of the interval as the only secondary node. Here, $f(1) = 1^2 = 1$, the area is 1, the whole square. We record this result as follows:

$$R_1 = 1\,,$$

where R stands for "right".





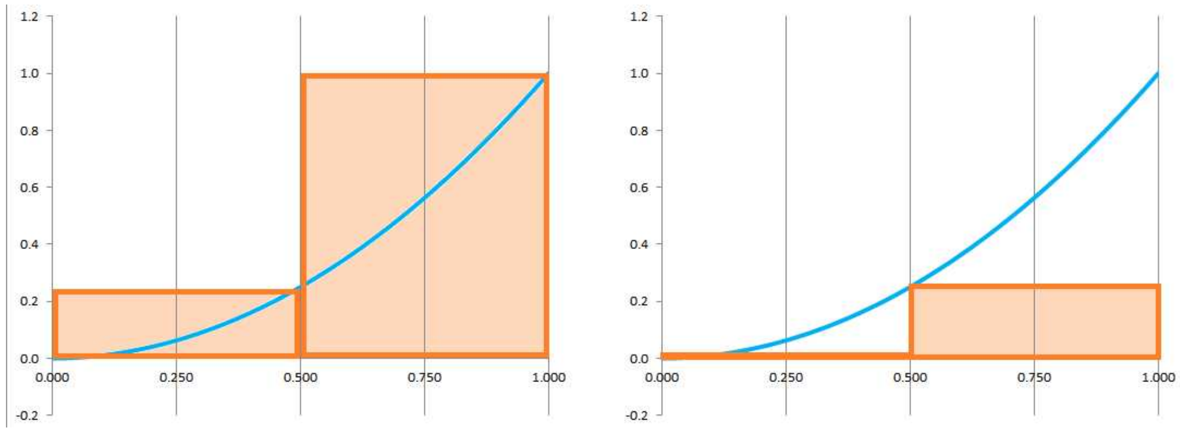
In the meantime, if we choose the *left* end, we have $f(1) = 0^2 = 0$, so the area is 0. We record this result as follows:

$$L_1 = 0\,,$$

where L stands for "left".

Next, $n = 2$. Then $\Delta x = 1/2$ and the interval is subdivided as follows:

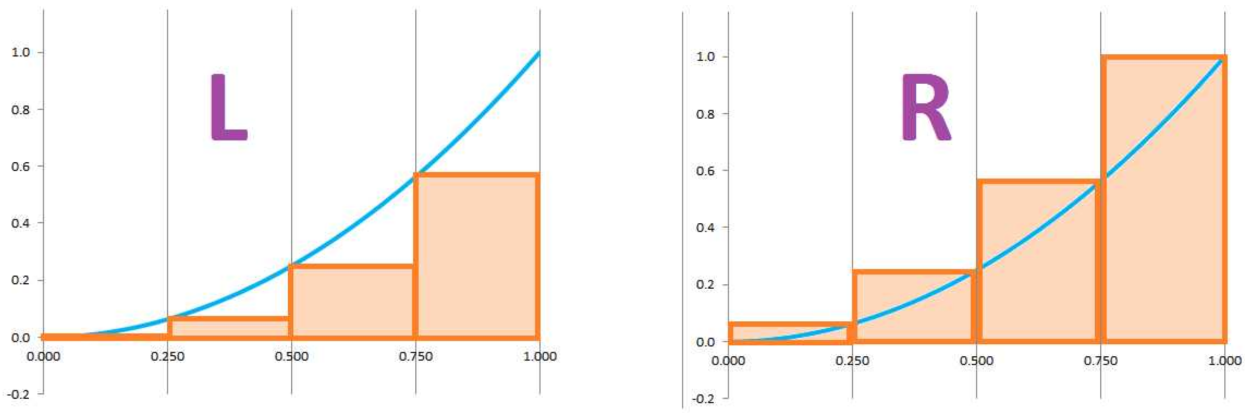
	$\bullet$	$---$	$\vert$	$---$	$\bullet$	
x	0		$1/2$		1	
$f(x) = x^2$	0		$1/4$			
L_2	$0 \cdot 1/2$	$+$	$1/4 \cdot 1/2$			$= 1/8$
R_2			$1/4 \cdot 1/2$	$+$	$1 \cdot 1/2$	$= 5/8$



Note that these two quantities underestimate and overestimate the area, respectively.

Now, $n = 4$. Then $\Delta x = 1/4$ and the interval is subdivided as follows:

	$\bullet$	$--$	$\vert$	$--$	$\vert$	$--$	$\vert$	$--$	$\bullet$	
x	0		$1/4$		$1/2$		$3/4$		1	
x^2	0		$1/16$		$1/4$		$9/16$		1	
L_4	$0 \cdot 1/4$	$+$	$1/16 \cdot 1/4$	$+$	$1/4 \cdot 1/4$	$+$	$9/16 \cdot 1/4$			≈ 0.22
R_4			$1/16 \cdot 1/4$	$+$	$1/4 \cdot 1/4$	$+$	$9/16 \cdot 1/4$	$+$	$1 \cdot 1/4$	≈ 0.47



Note that these two quantities underestimate and overestimate the area, respectively.

This is how the algebraic representation of the computation in the table:

$$\begin{aligned} R_4 &= \frac{1}{4}\cdot\frac{1}{16}+\frac{1}{4}\cdot\frac{1}{4}+\frac{1}{4}\cdot\frac{9}{16}+\frac{1}{4}\cdot 1 \\ &= \frac{1}{4}\left(\frac{1}{16}+\frac{4}{16}+\frac{9}{16}+\frac{16}{16}\right) \\ &= \frac{1}{4}\frac{30}{16} \\ &\approx 0.47\,. \end{aligned}$$

We continue with larger and larger values of n. We end up with a *sequence* of numbers. Then the limit of this sequence is meant to produce the area of this curved figure. These expressions are the *Riemann sums* of the function. This how it works:

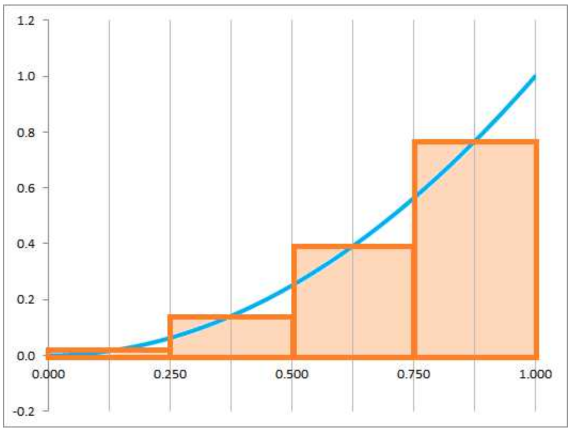
- The Riemann sums *approximate* the exact area.
- The limit of the Riemann sums *is* the exact area.

Example 1.5.2: mid-points

There are other Riemann sums if we choose other secondary nodes of the intervals, such as mid-points; it is denoted by M_n. Let's again estimate the area under $y = x^2$ with $n = 4$. Since L_4 and R_4 underestimate or overestimate the area respectively, one might expect that M_4 will be closer to the truth.

The value of Δx is still $1/4$:

	$\bullet \quad --$	$\vert$	$--$	$\vert$	$--$	$\vert$	$--$	$\bullet$
x	$1/8$		$3/8$		$5/8$		$7/8$	
$f(x) = x^2$	$(1/8)^2$		$(3/8)^2$		$(5/8)^2$		$(7/8)^2$	
M_4	$(1/8)^2 \cdot 1/4$	$+$	$(3/8)^2 \cdot 1/4$	$+$	$(5/8)^2 \cdot 1/4$	$+$	$(7/8)^2 \cdot 1/4$	≈ 0.328



Exercise 1.5.3

Finish the sentence "L_n and R_n will continue to underestimate and overestimate the area no matter how many intervals we have because f is..."

Exercise 1.5.4

Finish the sentence "M_n will continue to underestimate the area no matter how many intervals we have because f is..."

Example 1.5.5: area of triangle

Let's test the Riemann sum approach to computing areas to another familiar region, a triangle. Suppose

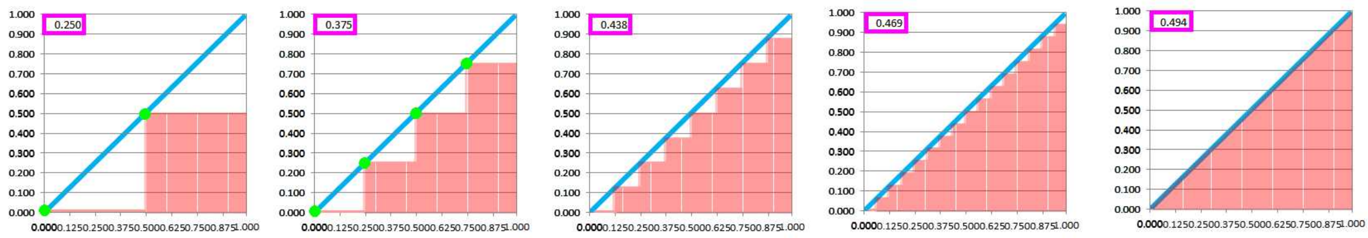
$$f(x) = x\,.$$

What is the area under its graph from $x = 0$ to $x = 1$?

We start with a left-end partition and these secondary nodes:

$$\begin{array}{ccccccccc} 0 & \rule{3em}{0.6em} & \frac{1}{n} & \rule{3em}{0.6em} & \frac{2}{n} & \ldots & \frac{n-1}{n} & \rule{3em}{0.6em} & 1 \\ 0 = c_1 & \rule{3em}{0.6em} & c_2 & \rule{3em}{0.6em} & c_3 & \ldots & c_n & \rule{3em}{0.6em} & \circ \end{array}$$

We next plot this chart for $n = 2\,,\ 4\,,\ 8\,,\ 16\,,\ 80$ to visualize the Riemann sums:



The last one has virtually no gaps. We let the spreadsheet to find the total areas (corners):

n	2	4	8	16	80	...	n
Δx	$1/2$	$1/4$	$1/8$	$1/16$	$1/80$	...	$1/n$
D_n	0.250	0.375	0.438	0.469	0.494	...	?

The numbers seem to approach 0.5 as expected. To find the full truth, we will need the *limit* of this sequence!

For an arbitrary n, the total area is only written recursively. Let's try to simplify $(f(x) = x)$:

$$\begin{array}{rcccccccl} D_n & = & f(c_1)\cdot\Delta x & + & f(c_2)\cdot\Delta x & + \ ... \ + & f(c_n)\cdot\Delta x & & \text{Substitute.} \\ & = & \frac{1}{n}\cdot\frac{1}{n} & + & \frac{2}{n}\cdot\frac{1}{n} & + \ ... \ + & \frac{n-1}{n}\cdot\frac{1}{n} & & \text{Factor it.} \\ & = & \Big(\quad 1 & + & 2 & + \ ... \ + & (n-1) & \Big)\cdot\frac{1}{n^2}\,. & \end{array}$$

How do we find this limit? The limit of the sum in parentheses is infinite and so is that of n^2. We have an indeterminate expression: $\infty \cdot 0$. Just as before, the only way to resolve it is *algebra*. We need to find an explicit formula for the sum.

Fortunately, the formula for the expression in parentheses is known (Chapter 1PC-1) as the sum of an arithmetic progression:

$$1 + 2 + ... + (n-1) = \frac{n(n-1)}{2}\,.$$

Therefore, the total area of the bars is the following limit as $n \to \infty$:

$$D_n = \frac{n(n-1)}{2}\frac{1}{n^2} = \frac{1}{2}\frac{n^2-n}{n^2} \to \frac{1}{2}\,,$$

according to the theorem *Limits of Rational Functions* (Chapter 2DC-2). The result matches what we know from geometry.

It should be clear that we wouldn't be able to apply the rules and methods of computing limits without this simplification step. Unlike most sequences we saw in differential calculus (Volume 2, Chapter 2DC-1), the expression for the Riemann sum has n terms. This means that we do not have a *direct*, or explicit, formula for the nth term of this sequence. Converting the former into the latter requires some challenging algebra. A few such formulas are known.

Theorem 1.5.6: Formulas for Finite Sums

The sums of m consecutive numbers, their squares, and their cubes are the

following:

$$\begin{aligned}
\sum_{k=1}^{m} k &= \frac{m(m+1)}{2} \\
\sum_{k=1}^{m} k^2 &= \frac{m(m+1)(2m+1)}{6} &&= \frac{m^3}{3}+\frac{m^2}{2}+\frac{m}{6} \\
\sum_{k=1}^{m} k^3 &= \left(\frac{m(m+1)}{2}\right)^2 &&= \frac{m^4}{4}+\frac{m^3}{2}+\frac{m^2}{4}
\end{aligned}$$

Proof.

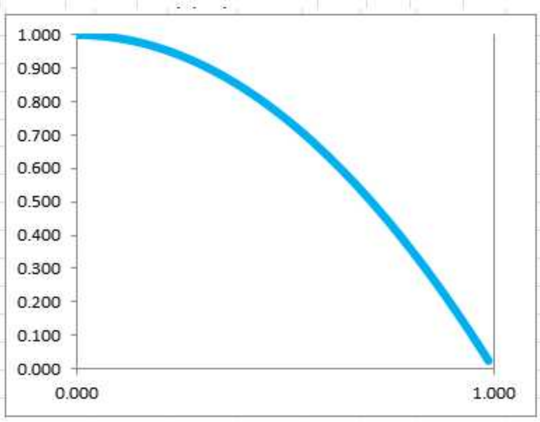
The first one is proven in Chapter 1PC-1.

Example 1.5.7: area under parabola

Problem: Suppose we are following a landing module on its trip to the moon. The very last part of the trip is very close to the surface and it is supposed to cover $2/3$ of a mile in one minute. Considering that there is no way to measure altitude above the surface, how would we know that it has landed?

This is what we do know: The velocity, at all times:

	time	velocity
1	0	1.000
2	0.01	1.000
3	0.03	0.999
4	0.04	0.999
5	0.05	0.998
6	0.06	0.996
7	0.08	0.994
8	0.09	0.992
9	0.1	0.990
10	0.11	0.987
11	0.13	0.984
12	0.14	0.981
13	0.15	0.978
14	0.16	0.974



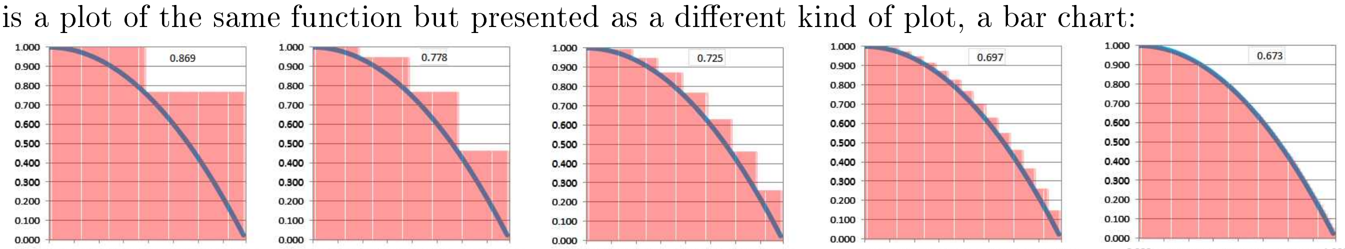
Solution: We approach the problem by predicting the displacement every:

- 1 minute, or
- $1/2$ minute, or
- $1/4$ minute, or
- etc.,

by using the speed recorded at the beginning of each of these intervals:

$$\text{projected displacement} = \text{initial velocity} \cdot \text{time passed}.$$

We have already plotted the velocity and now we are also to visualize the projected displacement. It is a plot of the same function but presented as a different kind of plot, a bar chart:



We plot this chart for different number of time intervals, n, and different choices of their lengths, $\Delta x = 1/n$. We also let the spreadsheet add these numbers to produce the total displacement, D_n,

over the whole one-minute interval. This is the data:

n	2	4	8	16	80
Δx	1/2	1/4	1/8	1/16	1/80
D_n	0.869	0.778	0.725	0.697	0.673

The data suggests that the craft has landed as the estimated displacement seems to be close to 2/3 miles. How close are we to this distance? As this question means different things to different people, let's try to find a *rule* for answering it: Find the projected displacement D_n as it depends on n.

First, the length of each time interval is:

$$\Delta x = \frac{1}{n}\,,$$

and the moments of time for sampling are:

$$x_1 = 0,\ x_2 = \frac{2}{n},\ x_3 = \frac{3}{n},\ ...,\ x_n = \frac{n-1}{n}\,.$$

Furthermore, we need exact, complete data about the velocity. Suppose the velocity y as a function of time, x, is given by this, exact formula:

$$y = f(x) = 1 - x^2\,.$$

Then, the total displacement is

$$\begin{array}{lllllll}
D_n & = & f(x_1)\cdot\Delta x & + & f(x_2)\cdot\Delta x & + \ ... \ + & f(x_n)\cdot\Delta x \\
& = & \left(1-\left(\frac{1}{n}\right)^2\right)\cdot\frac{1}{n} & + & \left(1-\left(\frac{2}{n}\right)^2\right)\cdot\frac{1}{n} & + \ ... \ + & \left(1-\left(\frac{n-1}{n}\right)^2\right)\cdot\frac{1}{n} \\
& = 1 - \Bigg(& \frac{1^2}{n^2} & + & \frac{2^2}{n^2} & + \ ... \ + & \frac{(n-1)^2}{n^2} \qquad \Bigg)\cdot\frac{1}{n} \\
& = 1 - \Bigg(& 1^2 & + & 2^2 & + \ ... \ + & (n-1)^2 \qquad \Bigg)\cdot\frac{1}{n^3}\,.
\end{array}$$

We are in the same place as in the last example: The sequence of improving approximations is the sequence of sums of a certain sequence. And, once again, the recursive expression must be converted to an explicit formula in order to apply the methods of computing limits that we know.

Let's simplify it using the last theorem:

$$1^2 + 2^2 + ... + (n-1)^2 = \frac{(n-1)^3}{3} + \frac{(n-1)^2}{2} + \frac{n-1}{6}\,.$$

Therefore,

$$\begin{aligned}
D_n &= 1 - \left(\frac{(n-1)^3}{3} + \frac{(n-1)^2}{2} + \frac{n-1}{6}\right)\frac{1}{n^3} \\
&= 1 - \frac{(n-1)^3}{3n^3} - \frac{(n-1)^2}{2n^3} - \frac{n-1}{6n^3}.
\end{aligned}$$

With this formula, we can improve the accuracy of our approximation of the total displacement to any degree we desire by choosing larger and larger values of n.

Next, what is the *exact* displacement D? It is the limit of the sequence D_n. We apply the theorem *Limits of Rational Functions* (Chapter 2DC-1)) and compare the leading terms of the two fractions:

$$D_n \to 1 - \frac{1}{3} - 0 - 0 = \frac{2}{3}.$$

Exercise 1.5.8

Redo the example for $f(x) = x^2$.

Exercise 1.5.9

Use the last formula in the theorem to find the area under the graph of $y = x^3$.

In order to escape the need for theorems like the one above, we will need to develop a work-around. An indirect approach will prove more effective.

1.6. The limit of the Riemann sums: the Riemann integral

We improve our approximations to produce better and estimates of the areas or displacements by making the partition finer and finer. We then consider the limit of this process.

We are, in fact, following the route of the difference quotient:

$$\begin{array}{ccc} \text{sequence} & \xleftarrow{\text{sampling}} & \text{function on } [a,b] \\ \downarrow{\scriptstyle \text{DQ}} & & \downarrow{\scriptstyle \text{derivative}} \\ \text{sequence} & \xrightarrow{\Delta x \to 0} & \text{function on } [a,b] \end{array} \qquad \begin{array}{ccc} \text{sequence} & \xleftarrow{\text{sampling}} & \text{function on } [a,b] \\ \downarrow{\scriptstyle \text{RS}} & & \downarrow{\scriptstyle \text{integral}} \\ \text{sequence} & \xrightarrow{\Delta x \to 0} & \text{function on } [a,b] \end{array}$$

Let's review the setup.

Suppose this time that a function f is defined at *all* points of interval $[a, b]$.

We are also considering *all* possible augmented partitions P of $[a, b]$:

$$a = x_0 \le c_1 \le x_1 \le c_2 \le x_2 \le \ldots < x_{n-1} \le c_1 \le x_n = b\,.$$

Then each such partition will produce the difference quotient of the function:

$$\frac{\Delta f}{\Delta x} = \frac{f(c_{k+1}) - f(c_k)}{\Delta x_k},$$

and the Riemann sum of the function:

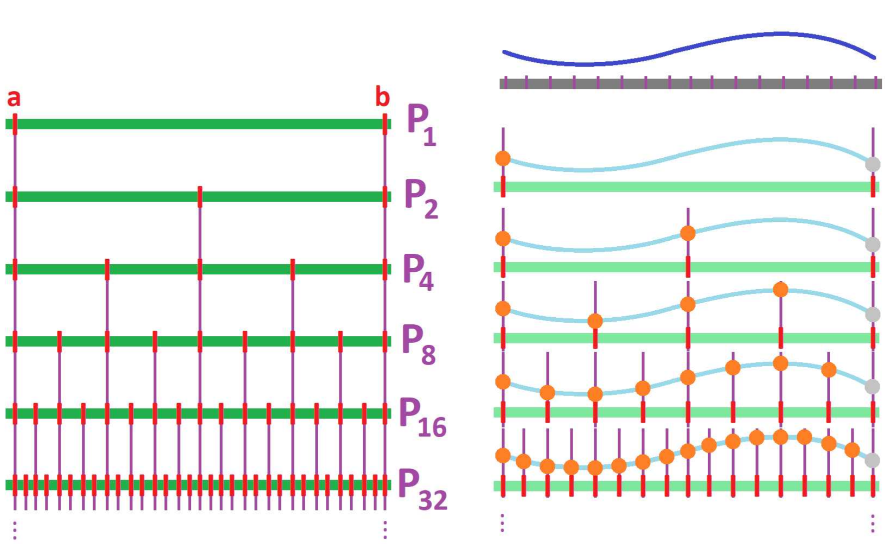
$$\Sigma f \cdot \Delta x = f(c_1)\,\Delta x_1 + f(c_2)\,\Delta x_2 + ... + f(c_n)\,\Delta x_n = \sum_{i=1}^{n} f(c_i)\,\Delta x_i\,,$$

where

$$\Delta x_i = x_i - x_{i-1}\,.$$

Now, in order to improve these approximations, we *refine* the partition: There will be more intervals and they are smaller. We keep refining. The result is a *sequence of partitions*, P_n.

Here is an example of a typical sequence of refining partitions. We simply cut every interval in half every time (left) again and again:



The left ends are chosen as secondary nodes to sample the function (right).

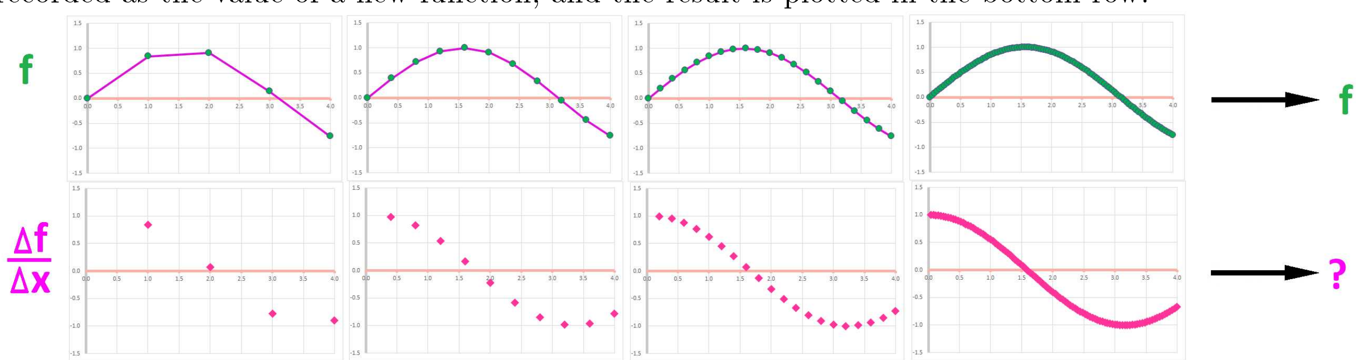
The result is what we can think of as a new function f_n.

Example 1.6.1: $\sin x$

Let's consider

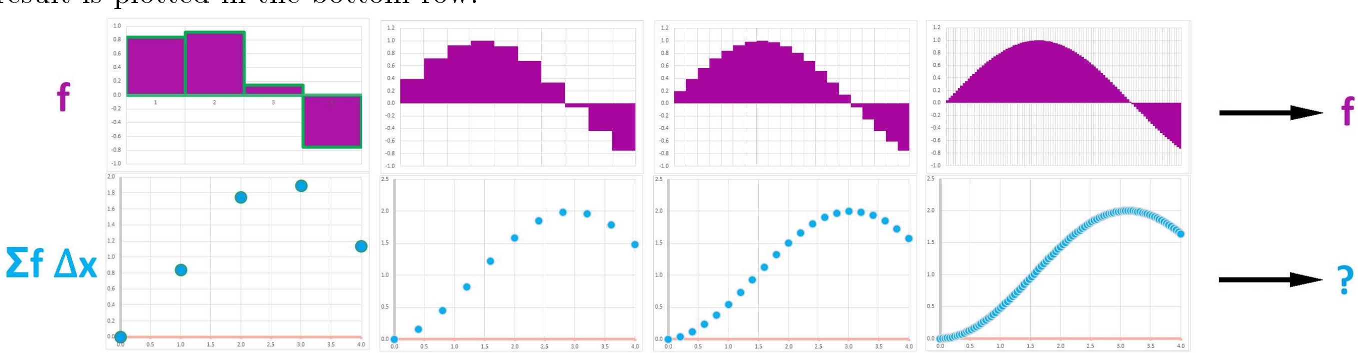
$$f(x) = \sin x\,.$$

For each $n = 2, 3, 4, \ldots$, the increment is found, $\Delta x = (b-a)/n$, and we have n segments in our partition of the interval. On each of its segments, the difference quotient is computed, the value is recorded as the value of a new function, and the result is plotted in the bottom row:



We face a *sequence of functions*. Its limit is the derivative.

Similarly the Riemann sum is computed, the value is recorded as the value of a new function, and the result is plotted in the bottom row:



We face a *sequence of functions* again. What is its limit?

Our sampling of the function is getting denser and denser. In the meantime, the points that make up the graphs of the difference quotients and those that make up the Riemann sums are getting closer and closer together. What is at the end of this process? A new function defined on the whole interval!

For simplicity, below we consider the Riemann sums on a *fixed interval*. We will be using $[a, b]$ instead of all

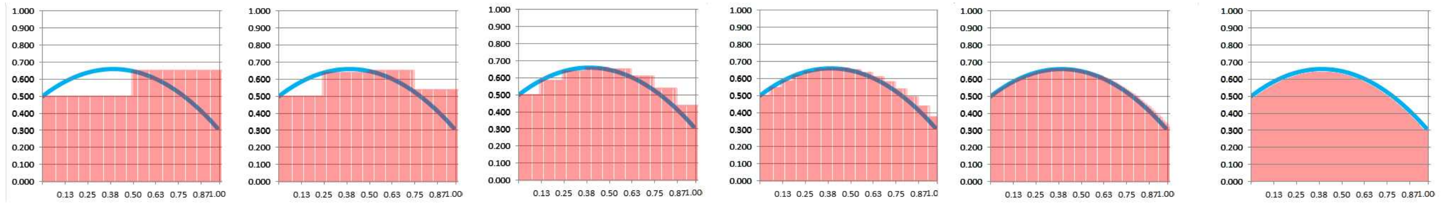
of these: $[a, x_k]$.

The Riemann sums over $[a, b]$ of the functions f_n acquired from sampling,

$$S_n = \Sigma f_n \Delta x \,,$$

form a *numerical sequence*, S_n.

As we have seen, this sequence may converge as $n \to \infty$:



The limit of this sequence, $\lim\limits_{n\to\infty} S_n = S$, is what we are after.

However, there are left-end, right-end, mid-point partitions, partitions with unequal increments, and infinitely many other sequences of partitions. Just as one can approximate a circle with non-regular polygons! We, therefore, need to consider *all* of them and require that *all* of them produce the same number:

$$\begin{array}{ccccc}
L_n & & M_n & & R_n \\
 & \searrow & \downarrow & \swarrow & \\
S_n & \to & {\color{red}S} & \leftarrow & ? \\
 & \nearrow & \uparrow & \nwarrow & \\
? & & ? & & ?
\end{array}$$

The only restriction is that each of these sequences of partitions have to be chosen in such a way that

$$\Delta x_i \to 0 \quad \text{as} \quad n \to \infty \,.$$

To make sense of this, we define the following:

Definition 1.6.2: mesh of partition

The *mesh of a partition* P is the maximal value of its increment:

$$|P| = \max_i \, \Delta x_i \,.$$

It is a measure of the degree of "refinement" of P. Whenever it goes to 0, all increments go to 0 too. We use it as follows:

Definition 1.6.3: Riemann integral

The *Riemann integral*, or the *definite integral*, of a function f over interval $[a, b]$ is defined to be the limit I of a sequence of its Riemann sums S_n over augmented partitions P_n with their mesh approaching 0 as $n \to \infty$; i.e.,

$$S_n \to \ I, \quad \text{provided} \quad |P_n| \to 0 \,,$$

when all these limits are equal to each other. When this limit is a number, f is called an *integrable function over* $[a, b]$. When all these limits are equal to $+\infty$ (or $-\infty$), we say that the integral is *infinite*.

The notation is similar to that for anti-derivatives:

Integral

$$\int_a^b f\, dx$$

It reads "the integral of f from a to b".

Abbreviated, the definition is written as follows:

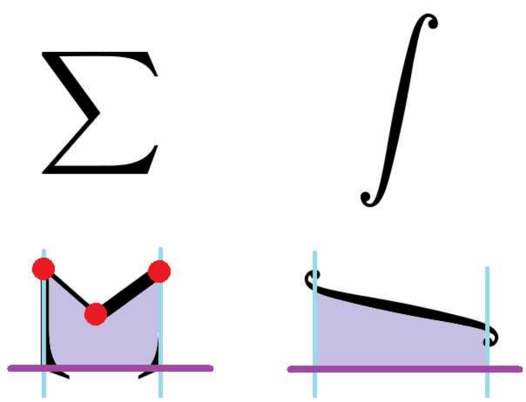
$$\int_a^b f\, dx = \lim_{n\to\infty} \Sigma f_n \Delta x\,,$$

where f_n is f sampled over the partition P_n.

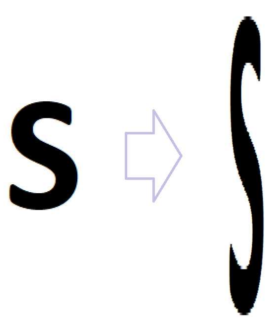
In the infinite case, we simply write (just as with other limits):

$$\int_a^b f\, dx = +\infty \ (\text{or } -\infty)\,.$$

So, the Riemann integral is the limit of the Riemann sums, in this special sense. Both can be seen as the areas under certain graphs:



The symbol "$\int$" is called the *integral sign*. It looks like a stretched letter S, which stands for "summation" (as does letter Σ):



The notation is similar, because it is related, to that for Riemann sums (it is also similar, because it is related, to that for antiderivatives). This is how the notation is deconstructed:

Riemann sum

left and right bounds for x

$\downarrow$

$$\text{domain}\left[\begin{array}{c} 1 \\ \\ -1 \end{array}\right. \int \left(3x^3 \qquad + \sin x \right) dx \;=\; 0$$

$\uparrow$ $\uparrow$ $\uparrow$

left and right brackets a specific number

We refer to a and b as the "lower bound" and the "upper bound" of the integral, while f is the "integrand". The interval $[a, b]$ is referred to as the "domain of integration".

Compare to the sigma notation:

left and right bounds for k

$\downarrow$

$$\text{domain}\left[\begin{array}{c} 10 \\ \\ k=0 \end{array}\right. \sum \left(3k^3 \qquad + \sin k \right) \;=\; 9076.411188$$

$\uparrow$ $\uparrow$ $\uparrow$

left and right brackets a specific number

Either the integral sign or the sigma sign designates a certain function that takes a function – of x or k – as its input. The output is a number.

Warning!

These "bounds" a and b aren't just lower and upper bounds of the interval $[a, b]$ but its minimum and maximum. It is also very common to use "limits" instead of "bounds".

Exercise 1.6.4

What do those "bounds" have to do with the word "boundary"?

While dx seems to be nothing but a "bookend" in the above notation, let's not forget that this is the *differential* (Chapter 2DC-3). An alternative notation reflects the fact that the integral is determined by the *differential form* $f(x) \cdot dx$ (Chapter 4):

Riemann integral

$$\int_{[a,b]} f\, dx$$

So, this is a function created by the differential form and its inputs are intervals. To emphasize the latter point, we move the *domain of integration* $[a, b]$ to the subscript.

This is how the notation is deconstructed:

Riemann integral

$$\begin{array}{ccccc} & \text{"evaluate"} & & & \\ & \downarrow & & & \\ & \int & \underbrace{(3x^3+\sin x)\,dx} & = & 0 \\ \text{input} \to & [-1,1] & \uparrow & & \uparrow \\ & & \text{differential form} & & \text{output} \end{array}$$

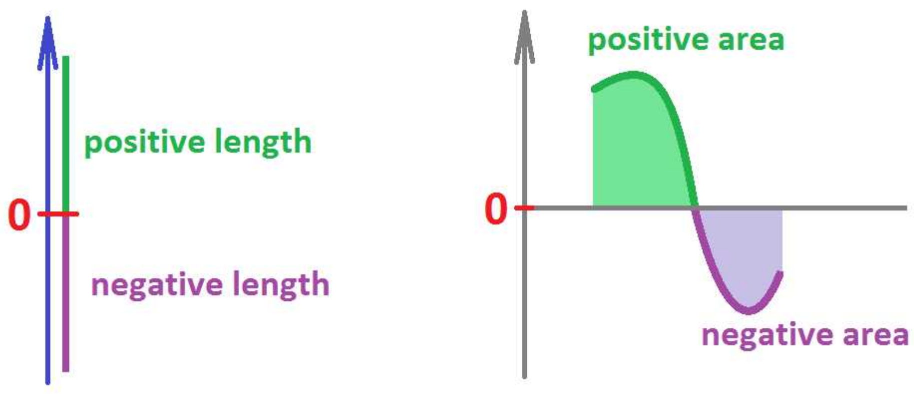
This transition becomes inevitable in higher dimensions (Volume 4).

When we speak of the area, we have (with n dependent on k):

$$\underbrace{\int_a^b f\,dx}_{\text{the exact area under the graph}} = \lim_{k\to\infty} \underbrace{\sum_{i=1}^n f(c_i)\,\Delta x_i}_{\text{areas of the bars}}\,.$$

Example 1.6.5: negative area?

Just as with the Riemann sums, when the values of f are negative, so is the "area" under its graph:



This demonstrates a drawback of the interpretation of the integral as the area in contrast to the interpretation as the displacement.

Let's verify that the definition makes sense for the simplest function.

Theorem 1.6.6: Integral of Constant Function

Suppose f is constant on $[a,b]$, i.e., $f(x)=c$ for all x in $[a,b]$ and some real number c. Then f is integrable on $[a,b]$ and we have:

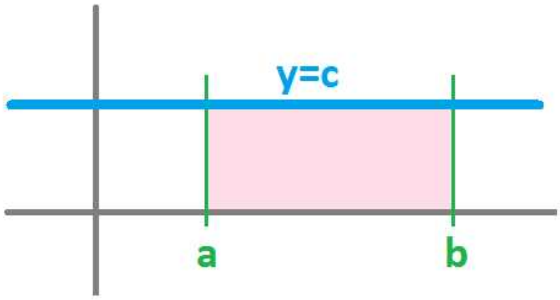
$$\boxed{\int_a^b f\,dx = c(b-a)}$$

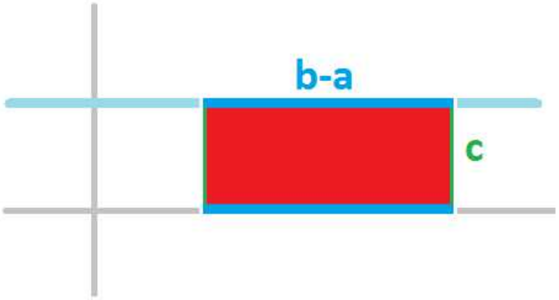
Proof.

From the *Constant Function Rule for Riemann Sums*, we know:

$$\Sigma_{[a,b]} f_k\,\Delta x = c(b-a)\,.$$

Since this expression doesn't depend on k, the conclusion follows.





Of course, we have recovered the formula for the area of a rectangle:

$$\text{area} \;=\; \text{height} \;\cdot\; \text{width}\,.$$

Except, this area is negative when c is negative.

Even though the answer to the question "How do we do it?" is still to come, we ask "Is it always possible?"

Some limits don't exist.

Then, as a limit, the integral might not exist either.

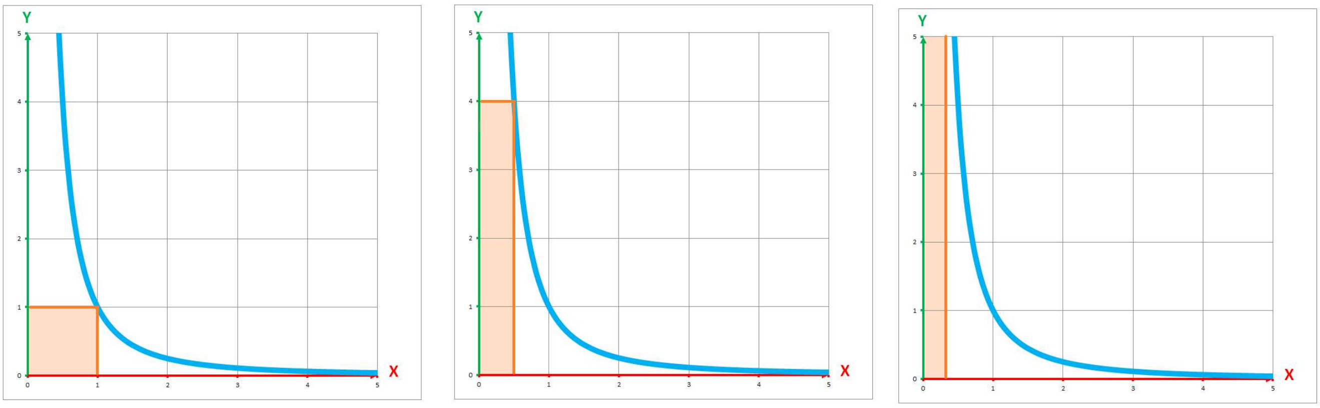
In contrast to differentiability, we can't tell by just looking at the graph.

Example 1.6.7: infinite area

Here is a simple example of a function non-integrable over $[0,1]$:

$$f(x)=\begin{cases}\dfrac{1}{x^2} & \text{if } x>0\,,\\ 0 & \text{if } x=0\,.\end{cases}$$

It suffices to look at the first term of the nth right-end Riemann sums:



This bar is getting taller and thinner at the same time. What about its area? We compute:

$$f(x_1)\,\Delta x = f\left(\frac{1}{n}\right)\frac{1}{n} = \frac{1}{1/n^2}\cdot\frac{1}{n} = n \to \infty \qquad \text{as } n\to\infty\,.$$

It turns out this bar is getting taller faster than it is getting thinner! Then, from the *Push Out Theorem for Divergence* (Chapter 2DC-1), it follows that

$$R_n \to \infty \quad \text{as } n\to\infty\,.$$

Therefore,

$$\int_0^1 f\,dx = +\infty\,.$$

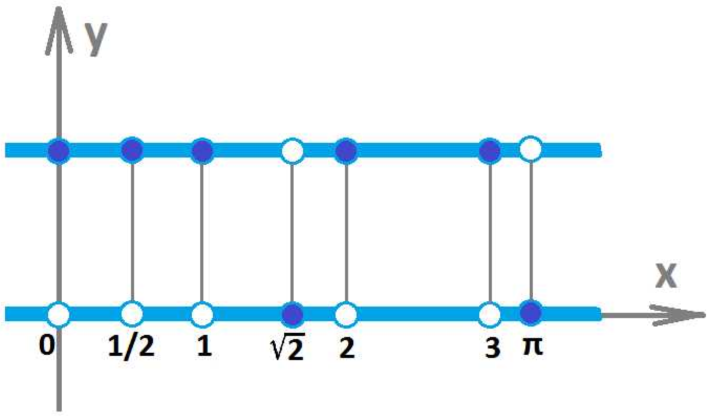
Exercise 1.6.8

Consider this construction for $\frac{1}{x}$ and $\frac{1}{x^3}$.

Example 1.6.9: Dirichlet function

The *Dirichlet function* is also non-integrable:

$$I_Q(x) = \begin{cases} 1 & \text{if } x \text{ is a rational number,} \\ 0 & \text{if } x \text{ is an irrational number.} \end{cases}$$



To prove this, we consider two different sequences of partitions:

- If we choose all the secondary nodes of partitions P_n to be rational, the corresponding Riemann sums will be equal to 1; therefore, the limit will be equal to $b - a$.
- If we choose all the secondary nodes of partitions Q_n to be irrational, the corresponding Riemann sums will be equal to 0; therefore, the limit is also equal to 0.

The mismatch between the limits of Riemann sums proves that the function is not integrable.

The following result proves that our definition makes sense for a large class of functions.

Theorem 1.6.10: Integrability of Continuous Functions

All continuous functions on $[a, b]$ *are integrable on* $[a, b]$.

We accept it without proof.

The converse isn't true because of the following.

Example 1.6.11: sign function

The sign function, $f(x) = \text{sign}(x)$, has a very simple graph and, it appears, the area under it would "make sense":

Indeed, this function is integrable over any interval $[a, b]$.

There are two cases. When $b \leq 0$ or $a \geq 0$, the function is simply constant on this interval. When $a < 0 < b$, all terms of all Riemann sums are $-1 \cdot \Delta x_i$, or $1 \cdot \Delta x_i$, or 0 (at most two). Let's suppose that 0 isn't a node of any of the partitions P_k and, in fact, it is one of its secondary nodes, $0 = c_m$.

Then

$$\begin{aligned}\Sigma f \cdot \Delta x &= \sum_{i=1}^{m-1}(-1)\cdot\Delta x_i \quad +0\cdot\Delta x_m \qquad + \sum_{i=m+1}^{n} 1\cdot\Delta x_i\\ &= -\sum_{i=1}^{m-1}\Delta x_i \qquad\qquad\qquad\qquad + \sum_{i=m+1}^{n}\Delta x_i\\ &= -(x_m - a) \qquad\qquad\qquad\qquad +(b - x_{m+1})\\ &= a+b \qquad\quad -x_m - x_{m+1}\,.\end{aligned}$$

Then, as $|P_k| \to 0$, we have:

$$x_m \to 0 \ \text{ and } \ x_{m+1} \to 0\,.$$

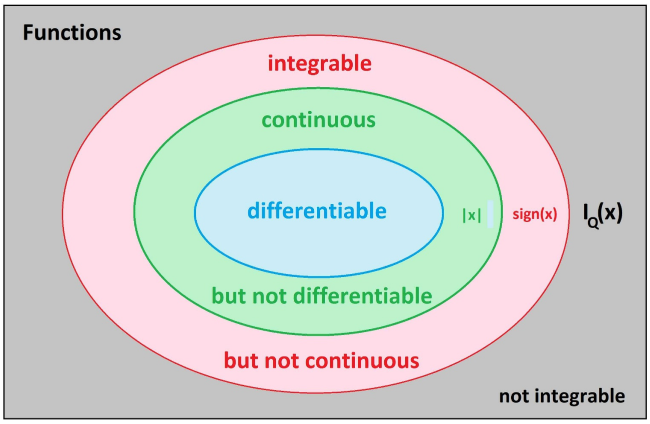
Therefore, the equation's limit is:

$$\int_a^b f\,dx = a+b\,.$$

Exercise 1.6.12

Consider the missing cases in the above proof.

These are the main classes of function we have seen and their relations:



They are *subsets* of each other!

We accept the following result without proof.

Theorem 1.6.13: Integrability of Restriction

If f is integrable over $[a,b]$, then it is also integrable over any $[A,B]$ with $a \le A < B \le b$.

Just as before, we observe that even if the person didn't spend any time driving, the displacement still makes sense; it's zero.

Theorem 1.6.14: Integral Over Zero-length Interval

The Riemann integral of a function f over a "zero-length" interval $[a,a]$, is equal to zero:

$$\boxed{\int_a^a f\,dx = 0}$$

And the area of a region one-point thick is zero, too.

Exercise 1.6.15

Prove the theorem.

In the differential form notation, this integral with equal bounds has a clearer meaning:

$$\boxed{\int_{\{a\}} f\,dx = 0}$$

The oriented intervals are also included.

We once again utilize the idea of oriented intervals and oriented rectangles.

Theorem 1.6.16: Integral Over Negatively Oriented Interval

The Riemann integral of a function f over a negatively oriented interval $[b,a]$, $b > a$ is equal to the negative of the integral over $[a,b]$:

$$\boxed{\int_b^a f\,dx = -\int_a^b f\,dx}$$

Exercise 1.6.17

Prove the theorem.

In the differential form notation, this "flipping" of the bounds of integral has a more precise meaning. We think of the negatively oriented interval as the negative of the original:

$$[b,a] = -[a,b]\,.$$

Then the formula takes the following form:

$$\boxed{\int_{-[a,b]} f\,dx = -\int_{[a,b]} f\,dx}$$

Example 1.6.18: area vs. integral

Thus, we have explained the meaning of the area of a curved region: It is an integral. Conversely, is the integral an area? Yes, in a sense. It depends on the *units* of the variables. For example, when both x and y are measured in feet, the integral is indeed the area in the usual sense; it is measured in square feet. However, what if x and y are something else? We have already seen that x may be time and y the velocity; they are measured in seconds and feet per second respectively. Then the integral is measured in seconds $\cdot$ feet per second, i.e., feet. That can't be area... In fact, both x and y may be quantities of arbitrary nature; then the units of the integral might be: pound $\cdot$ degree, man-hour, etc. Then the integral is described as "the total value" of the function.

Here is the summary of the development of calculus up to this point:

Differential calculus	**Integral calculus**
f defined on the primary nodes	g defined on the secondary nodes
1. difference, Δf defined on the secondary nodes $\Delta f\,(c_k) = f(x_k) - f(x_{k-1})$	1. sum, Σg defined on the primary nodes $\Sigma g\,(x_k) = g(c_1) + ... + g(c_k)$
divide by Δx	Δx is factored in each term
2. difference quotient, $\dfrac{\Delta f}{\Delta x}$ defined on the secondary nodes $\dfrac{\Delta f}{\Delta x}(c_k) = \dfrac{f(x_k - f(x_{k-1})}{\Delta x}$	2. Riemann sum, $\Sigma g \cdot \Delta x$ defined on the primary nodes $\Sigma g \cdot \Delta x\,(x_k) = g(c_1)\Delta x + ... + g(c_k)\Delta x$
$\Delta x \to 0$	$\Delta x \to 0$
3. derivative, $\dfrac{df}{dx}$ defined on the interval	3. Riemann integral, $\displaystyle\int g\,dx$ defined on the interval

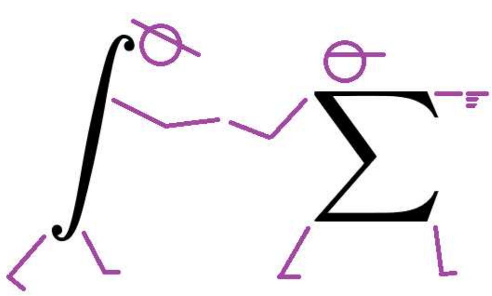
The two columns are constructed independently. Rows 1 and 2 are linked together by the discrete version of the Fundamental Theorem of Calculus. Row 3 will be linked together by its final version.

1.7. *Properties of the Riemann sums and the Riemann integrals*

We go back to the definitions in order to confirm that our theory makes sense by matching the perceived ideas of how these concepts are supposed to operate in real – though idealized – life. The main areas are, as before, motion and geometry.

The properties of the Riemann sums come from pure algebra. What's left, then, is to make sure that those relations are preserved under the *limit* to produce the matching properties of the Riemann integrals.

The integral follows the Riemann sum, every time:



The formula of the Riemann sum is:

$$\Sigma f \cdot \Delta x = f(c_1)\,\Delta x_1 + f(c_2)\,\Delta x_2 + ... + f(c_n)\,\Delta x_n\,,$$

where the points

$$a = x_0 \leq c_1 \leq x_1 \leq c_2 \leq x_2 \leq ... \leq c_n \leq x_n = b$$

make up an augmented partition of $[a, b]$.

The Riemann sum remains just a *sum*. We, therefore, can use some of the very elementary algebraic facts.

While adding, we can re-group the terms freely; in particular, we can remove the parentheses:

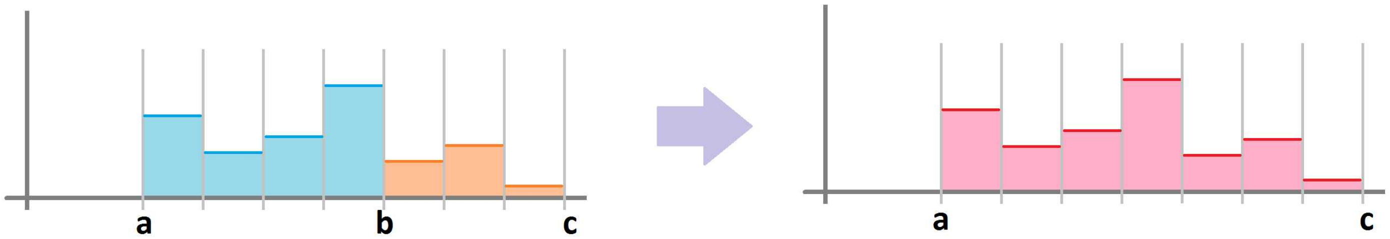
$$(a + b) + c = a + (b + c) = a + b + c\,.$$

For sums, we have:

$$(u_1 + u_2 + ... + u_n) + (v_1 + v_2 + ... + v_m) = u_1 + u_2 + ... + u_n + v_1 + v_2 + ... + v_m\,.$$

The statement is about the fact that when adding, we can change the order of terms freely; this is called the *Associativity Property* of addition. This is also the *Additivity Rule for Sums* (seen in Volume 1, Chapter 1PC-1).

Not much to change when we consider the sums of functions defined on a partition. If we have partitions of two adjacent intervals, we can just continue to add terms, thus creating a "longer" sum:



Or we think of these as areas.

The algebra is as follows:

Theorem 1.7.1: Additivity of Sums

The sum of the sums over two consecutive parts of an interval is the sum over the whole interval.

In other words, for any function f and for any function f and for any partitions of intervals $[a, b]$ and $[b, c]$, we have:

$$\boxed{\Sigma g\Big|_{[a,b]} + \Sigma g\Big|_{[b,c]} = \Sigma g\Big|_{[a,c]}}$$

What if these sums are *Riemann* sums? Not much changes:

Theorem 1.7.2: Additivity of Riemann Sums

The sum of the Riemann sums over two consecutive parts of an interval is the Riemann sum over the whole interval.

In other words, for any function f and for any partitions of intervals $[a, b]$ and $[b, c]$, we have:

$$\boxed{\Sigma f \cdot \Delta x\Big|_{[a,b]} + \Sigma f \cdot \Delta x\Big|_{[b,c]} = \Sigma f \Delta x\Big|_{[a,c]}}$$

Proof.

Suppose the two augmented partitions, P and Q, are given by:

$$P: \quad a = x_0 \le c_1 \le x_1 \le c_2 \le x_2 \le ... \le c_n \le x_n = b$$

$$Q: \quad b = y_0 \le d_1 \le y_1 \le d_2 \le y_2 \le ... \le d_m \le x_m = c$$

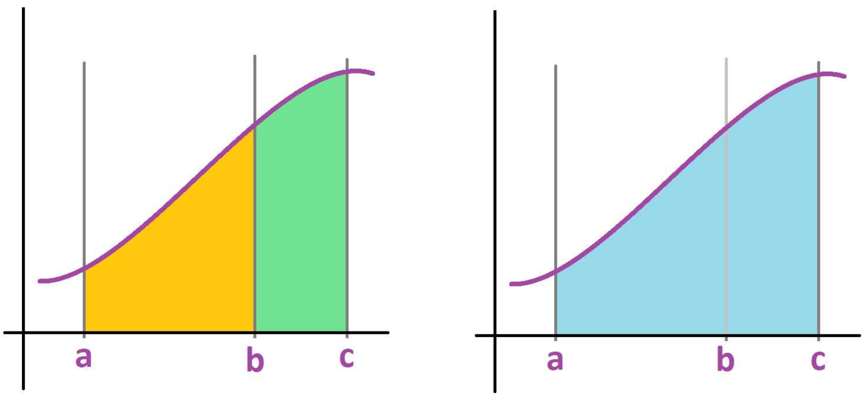
We rename the items on the latter list and form an augmented partition of $[a, c]$:

$$P \cup Q: \ a = x_0 \le c_1 \le x_1 \le c_2 \le ... \le c_n \le x_n \le c_{n+1} \le x_{n+1} \le c_{n+2} \le x_{n+2} \le ... \le c_{n+m} \le x_{n+m} = c$$

Applying the definition to the intervals $[a, b]$ and $[b, c]$ and then to $[a, c]$, we have:

$$\begin{aligned} \Sigma f \cdot \Delta x\Big|_{[a,b]} + \Sigma f \cdot \Delta x\Big|_{[b,c]} &= (f(c_1) + f(c_2) + ... + f(c_n)) \\ &\quad + (f(c_{n+1}) + f(c_{n+2}) + ... + f(c_{n+m})) \\ &= \Sigma f \Delta x\Big|_{[a,c]}. \end{aligned}$$

To use the *area* metaphor, imagine that we have zoomed out of the picture of the Riemann sums:



The result is equally applicable to the integrals; the formula in the theorem adds the areas of these two adjacent regions:

Theorem 1.7.3: Additivity of Integral

The sum of the integrals over two consecutive parts of an interval is the integral over the whole interval.

In other words, for any function f integrable over $[a, b]$ and over $[b, c]$ is also integrable $[a, c]$, and we have:

$$\boxed{\int_a^b f\,dx + \int_b^c f\,dx = \int_a^c f\,dx}$$

Proof.

The integrability follows from the last theorem in the last section. To prove the formula, we start with this fact about the mesh of partitions:

$$|P \cup Q| = \max\{|P|, |Q|\}.$$

Once we move to sequence of partitions, this fact implies the following:

$$|P_k \cup Q_k| \to 0 \iff |P_k| \to 0 \text{ and } |Q_k| \to 0\,.$$

Next, we take the formula in part (A), the Additivity of Riemann sums, take the limit with $k \to \infty$ and use the *Sum Rule for Limits* from Chapter 2DC-1.

As you can see, the word "additivity" in the name of the theorem doesn't refer to adding the terms of the Riemann sums but to *adding the domains of integration*, i.e., the union of the two intervals. The idea becomes especially vivid when the formula is written in the differential form notation:

$$\begin{array}{ccccc} \displaystyle\int_{[a,b]} f\,dx & + & \displaystyle\int_{[b,c]} f\,dx & = & \displaystyle\int_{[a,b]\cup[b,c]} f\,dx \\ \text{orange} & + & \text{green} & = & \text{blue} \end{array}$$

The interval becomes the input variable, subject to some algebra.

Exercise 1.7.4

Finish the formula:

$$\int_{[a,b]} f\,dx + \int_{[c,d]} f\,dx = \ldots$$

For the *motion* metaphor, we have:

$$\begin{array}{cl} & \text{distance covered during the 1st hour} \\ + & \text{distance covered during the 2nd hour} \\ = & \text{distance during the two hours} \end{array}$$

The following is an important corollary.

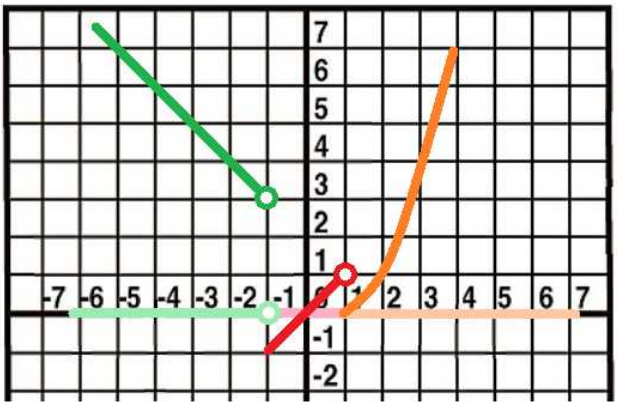
We also accept it without proof the corollary below.

Theorem 1.7.5: Integrability of Piecewise Continuous Functions

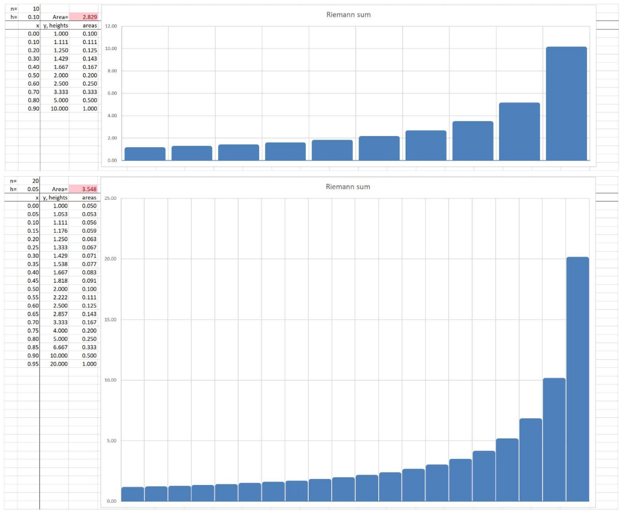
All piecewise continuous functions (i.e., continuous on all but a finite number of points with only jump discontinuities) on $[a,b]$ are integrable on $[a,b]$.

Proof.

It follows from the integrability of continuous functions (last section) and the *Additivity Rule*.



In particular, all step-functions are integrable. After all, they even look like Riemann sums:



As you can see, the properties of Riemann integrals follow from the corresponding properties of the Riemann sums – via the corresponding rules of limits. The graphical interpretations of the two properties are also the same; we just zoom out.

Suppose we have the right-end sum; there are n intervals of equal length,

$$\Delta x_i = h = \frac{b-a}{n},$$

between a and b and these are the secondary nodes:

$$c_i = \ a, \ a+h, \ \dots, \ b-h.$$

A proof for left-end sum and mid-points sums would be virtually identical.

Below, we will be reviewing some facts about *sums* presented in Chapter 1PC-1 and then applying them to the Riemann sums producing, via limits, results about the Riemann integrals.

Can we compare the values of two Riemann sums? Consider this simple algebra:

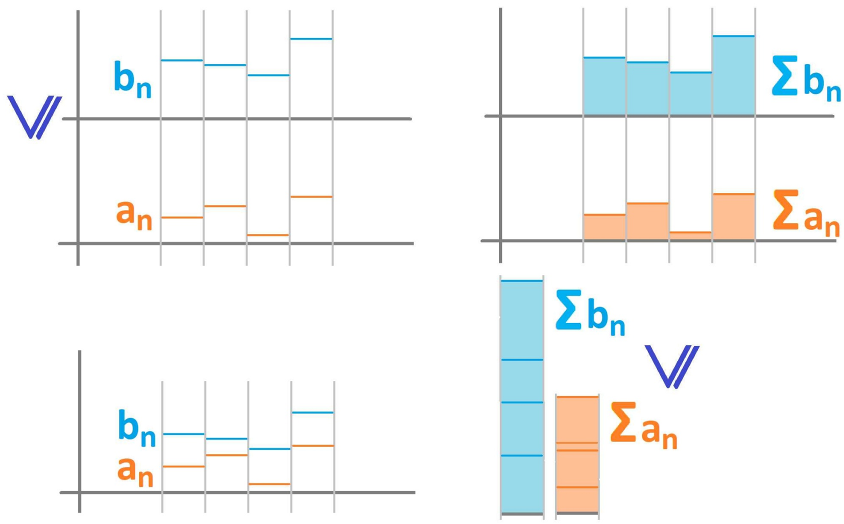
$$\begin{array}{lcl} u & \le & U \\ v & \le & V \\ \hline u+v & \le & U+V \end{array}$$

We can keep adding terms:

$$\begin{array}{rcl} u_p & \le & U_p, \\ u_{p+1} & \le & U_{p+1} \\ \vdots & \vdots & \vdots \\ u_q & \le & U_q \\ \hline u_p + \dots + u_q & \le & U_p + \dots + U_q \\ \displaystyle\sum_{n=p}^{q} u_n & \le & \displaystyle\sum_{n=p}^{q} U_n \end{array}$$

That's the *Comparison Rule for Sums*(seen in Volume 1, Chapter 1PC-1).

For functions defined over partitions, if one function "dominates" another, then so does its sum, Riemann sum, and later the integral:



We have the following:

Theorem 1.7.6: Comparison Rule for Sums

The sum of a smaller function is smaller.

In other words, if f and g are functions, then for any partition of an interval $[a,b]$, we have:

$$\boxed{f(x) \geq g(x) \text{ on } [a,b] \implies \Sigma f \ \geq \Sigma g}$$

where the summations are over an augmented partition of $[a,b]$.

What if these sums are *Riemann* sums?

Theorem 1.7.7: Comparison Rule for Riemann Sums

The Riemann sum of a smaller function is smaller.

In other words, if f and g are functions, then for any partition of an interval $[a,b]$, we have:

$$\boxed{f(x) \geq g(x) \text{ on } [a,b] \implies \Sigma f \cdot \Delta x \geq \Sigma g \cdot \Delta x}$$

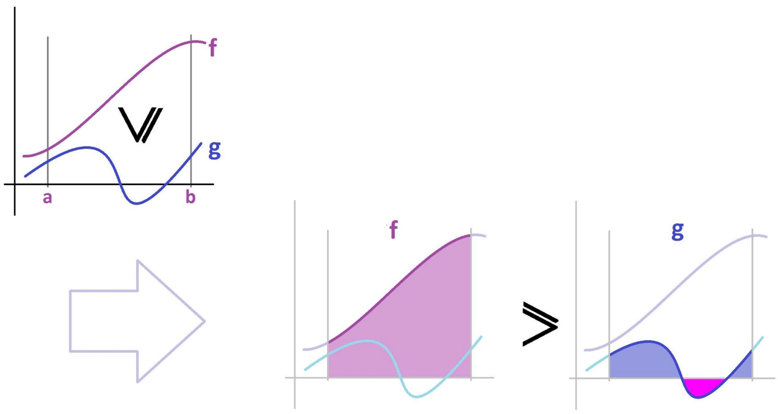
where the summations are over an augmented partition of $[a,b]$.

Proof.

Applying the definition of the Riemann sum to the function f and then to g, we have:

$$\begin{aligned}\Sigma f \cdot \Delta x\,(b) &= f(a) + f(a+h) + f(a+2h) + ... + f(b-h)\\ &\geq g(a) + g(a+h) + g(a+2h) + ... + g(b-h)\\ &= \Sigma g \cdot \Delta x\,(b)\,.\end{aligned}$$

If we zoom out, we see that the larger function always contains a larger area under its graph:



So, we take the limit of this inequality as $\Delta x \to 0$:

Theorem 1.7.8: Comparison Rule for Riemann Integrals

The integral of a smaller function is smaller.

In other words, if f and g are functions, then for any a, b with $a < b$, we have:

$$\boxed{f(x) \geq g(x) \text{ on } [a,b] \implies \int_a^b f\,dx \geq \int_a^b g\,dx}$$

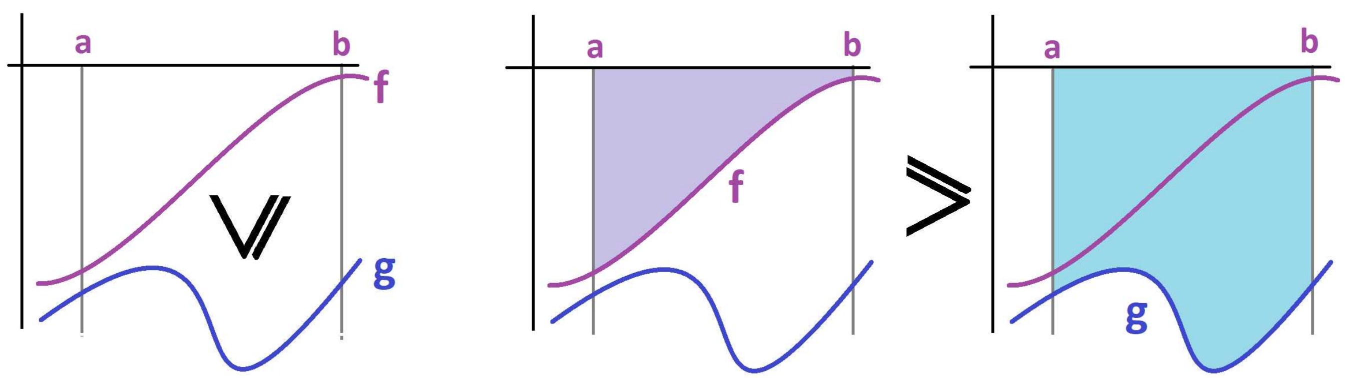
provided f and g are integrable over $[a,b]$.

Proof.

Now take the limit with $n \to \infty$ and use the *Comparison Rule for Limits* from Chapter 2DC-1.

Example 1.7.9: comparison of negative integrals

Unfortunately, the *area* interpretation of the Riemann integral will match our intuition only for as long as the areas are positive. The reason is that we measure the areas *from the x-axis to the graph.* That is why when the functions are negative, the areas are negative too. The comparison then appears to be wrong:



But the comparison hasn't been flipped!

There may be a similar conflict for the interpretation of this result in terms of *motion*. This simple interpretation fails because it doesn't take into account the directions: The faster covers the longer distance. Better still: The faster you go in a particular direction, the farther you progress.

Exercise 1.7.10

Prove the rest of the theorem.

Exercise 1.7.11

Modify the proof for: (a) left-end, (b) mid-point, and (c) general Riemann sums.

Exercise 1.7.12

What happens when $a > b$?

Related results are the following:

Theorem 1.7.13: Strict Comparison Rule for Sums

The sum of a strictly smaller function is strictly smaller.

In other words, if f and g are functions, then for any partition of an interval $[a, b]$ with $a < b$, we have:

$$\boxed{f(x) < g(x) \text{ on } [a, b] \implies \Sigma f < \Sigma g}$$

where the summations are over an augmented partition of $[a, b]$.

Theorem 1.7.14: Strict Comparison Rule for Riemann Sums

The Riemann sum of a strictly smaller function is strictly smaller.

In other words, if f and g are functions, then for any partition of an interval $[a, b]$ with $a < b$, we have:

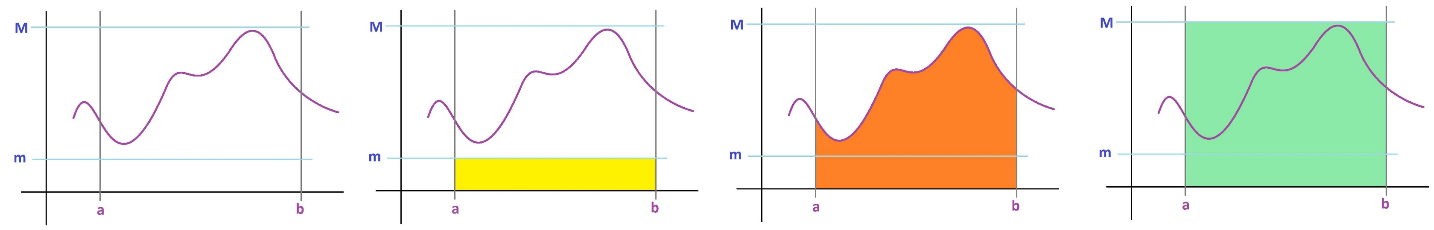
$$\boxed{f(x) < g(x) \text{ on } [a, b] \implies \Sigma f \cdot \Delta x < \Sigma g \cdot \Delta x}$$

where the summations are over an augmented partition of $[a, b]$.

Exercise 1.7.15

State and prove the Strict Comparison Rule for integrals. Hint: There is no Strict Comparison Rule for limits.

What if we know only *a priori* bounds of the function? Suppose the range of a function f lies within the interval $[m, M]$. Then its graph lies between the horizontal lines $y = m$ and $y = M$, with $m < M$.



Then, what can we say about the Riemann sum of f? We would like to find an estimate for the area of *the orange region* in terms of a, b, m, M. Below, the yellow region on the left is less than the orange area. On the right, the green area is larger.

Furthermore, these two regions are rectangles and their areas are easy to compute. We have the following inclusions:

$$\text{the smaller rectangle } \subset \text{ the region under the graph } \subset \text{ the larger rectangle}\,.$$

They prove the following inequalities:

$$\text{the area of the smaller rectangle} \ \le \ \text{the area under the graph} \le \ \text{the area of the larger rectangle}\,.$$

For the general case, we have the following:

Theorem 1.7.16: Bounds for Riemann Sums

Suppose f is a function and that for any partition of an interval $[a,b]$, we have:

$$m \le f(x) \le M\,,$$

for all x with $a \le x \le b$. Then, we have:

$$\boxed{m(b-a) \le \Sigma f \cdot \Delta x \le M(b-a)}$$

where the summation is over an augmented partition of $[a,b]$.

Proof.

For either this inequalities and the one in the next theorem, we apply the *Constant Function Rule* from the last section and the *Comparison Theorem* above. Now take the limit with $k \to \infty$ and use the *Bounds Rule for Limits* from Chapter 2DC-1.

Now we take the limit.

Theorem 1.7.17: Bounds for Integrals

Suppose f is a function and that for any a, b with $a < b$, we have:

$$m \le f(x) \le M\,,$$

for all x with $a \le x \le b$. If f is an integrable function over $[a,b]$, then we have:

$$\boxed{m(b-a) \le \int_a^b f\,dx \le M(b-a)}$$

In other words, the bounds for a function create bounds for the integral.

Exercise 1.7.18

State and prove the Strict Bounds for Integrals.

Example 1.7.19: a priori estimates

We know, *a priori*, that the values of such a trigonometric function as sin (and cos) lie within $[-1,1]$. Then, we have:

$$\begin{array}{rcccll} -1 & \le & \sin x & \le & 1 & \text{for all } x \quad \Longrightarrow \\ -1(b-a) & \le & \displaystyle\int_a^b \sin x\,dx & \le & 1(b-a) & \end{array}$$

So, even though all we know about the function is just these two, very crude, estimates, we *guarantee* that the integral will lie within the interval $[a-b, b-a]$.

1.8. The Fundamental Theorem of Calculus, continued

So far, we have learned the following:

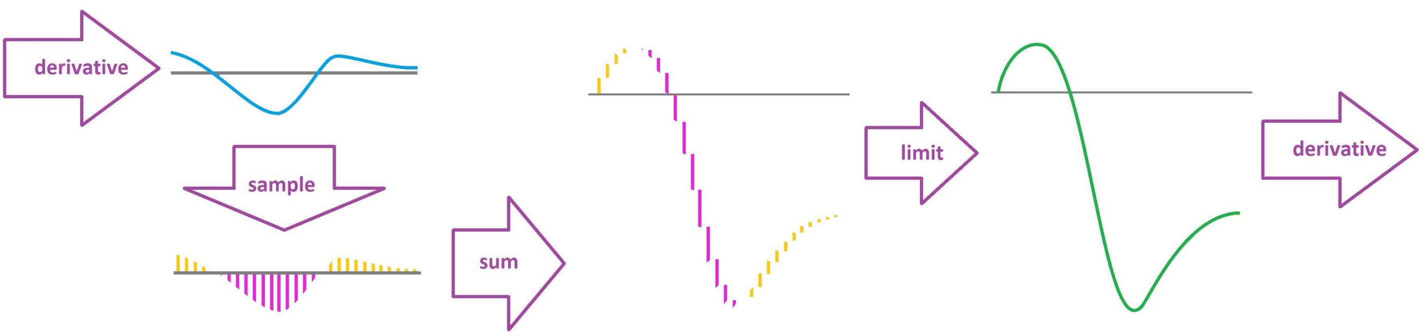
- ▶ The difference quotient and the Riemann sum undo the effect of each other.

The simple reason is that the algebraic operation they are created by cancel each other one by one.

Now, since the derivative and the Riemann integral are the *limits* of these two respectively, it stands to reason that they have the same relation:

- ▶ The derivative and the Riemann integral undo the effect of each other.

Below, the function is sampled to produce the Riemann sums, which under the limit produce the Riemann integral to be differentiated:



Will we make the full circle?

So, what does the Riemann integral have to do with antiderivatives? Same as what the Riemann sum does!

Example 1.8.1: positions and velocities

We already know the answer, for the *motion* metaphor. If $y = f(x)$ is the velocity and x is the time, then we have:

- One of the antiderivatives is the position function, $F(x)$.
- The integral is the displacement during $[a, b]$, $\int_a^b f\,dx$.

If we know the position at all times, we can certainly compute the displacement at any moment:

$$\underbrace{F(b)}_{\text{current position}} - \underbrace{F(a)}_{\text{initial position}}$$

Conversely, the displacement is found for any $x > a$ as $\int_a^x f\,dx$. After all, the two are attributes of the same motion, no matter how they are computed; they co-exist.

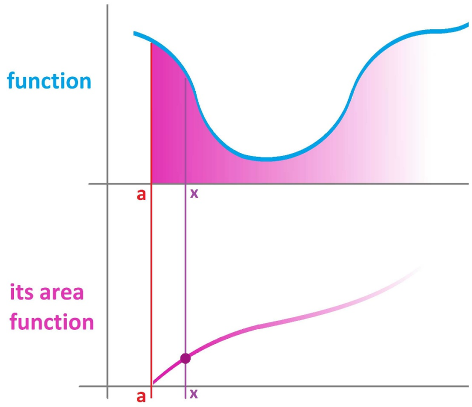
Example 1.8.2: areas and tangents

Now, what about the *area* metaphor? It's not transparent.

Analogous to what we did above for the Riemann sum, we define the *area function* of f to be:

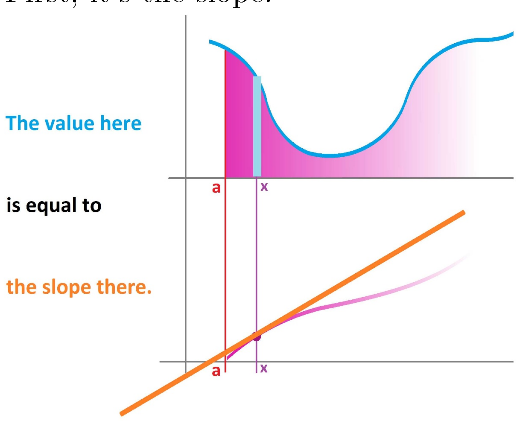
$$A(x) = \int_a^{\overbrace{x}^{\text{varies}}} f(t)\,dt = \text{ area under the graph over } [a, x]\,.$$

This Riemann integral has a *variable upper bound.* It's also illustrated below: x runs from a to b and beyond.



We can see how, wherever f has *positive values*, every increase of $x > a$ adds a slice to the area (less in the middle) causing A to *increase* (slower in the middle). Moreover, wherever f has *negative values*, this slice will have a negative area, thereby causing A to *decrease*. That's a behavior of an antiderivative of f!

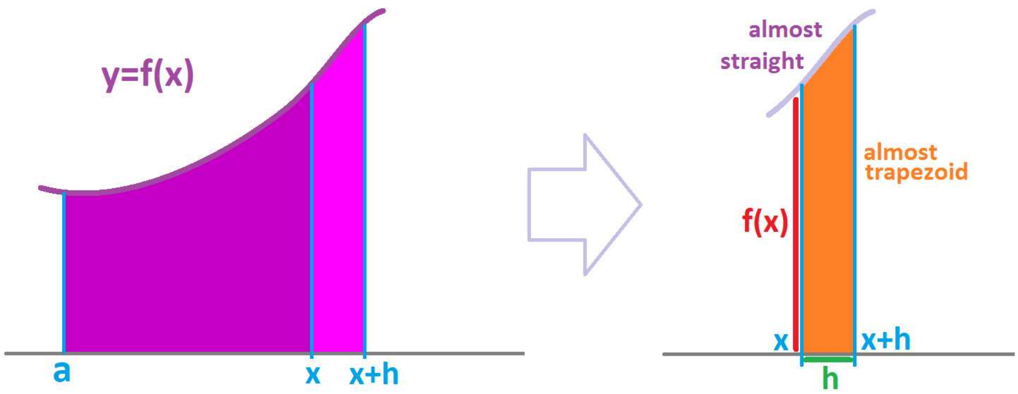
Let's consider its derivative, A'. First, it's the slope:



We go all the way back to the definition of the derivative:

$$\begin{aligned} A'(x) &= \lim_{h\to 0} \frac{A(x+h) - A(x)}{h} \\ &= \lim_{h\to 0} \frac{1}{h} \left(\int_a^{x+h} f\,dt - \int_a^x f\,dt \right) \\ &= \lim_{h\to 0} \frac{1}{h} \int_x^{x+h} f\,dx\,. && \text{According to the Additivity Rule.} \end{aligned}$$

The last value is illustrated on the right:



We are left with a slice of the area above the segment $[x, x+h]$; it looks like a trapezoid when h is small. What happens to this limit?

Let's make h small. The "trapezoid" will be thinner and thinner and its top edge will look more and more straight, assuming f is continuous. We know that the area of a trapezoid is the length of the mid-line times the height. Then we have:

$$\frac{\text{area of trapezoid}}{\text{width}} = \text{height in the middle}\,.$$

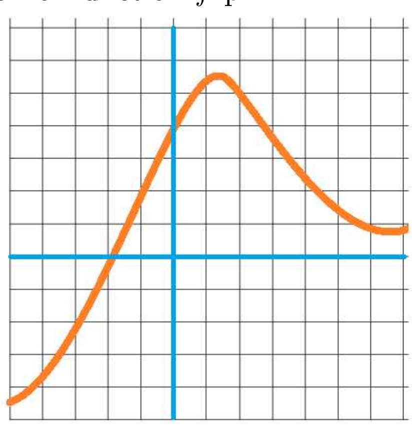
It follows that its height is approximated by

$$\frac{\int_x^{x+h} f\,dx}{h}\,.$$

This outcome suggests that $A' = f$. However, proving that the limit exists would require a subtler argument.

Exercise 1.8.3

Sketch the graph of the area function of function f plotted below for $a = 1$:



Exercise 1.8.4

Suppose the function plotted above is the area function of some function. Sketch the graph of that function.

Exercise 1.8.5

Prove that the area function is continuous.

Exercise 1.8.6

Finish the proof by using the *Squeeze Theorem* for:

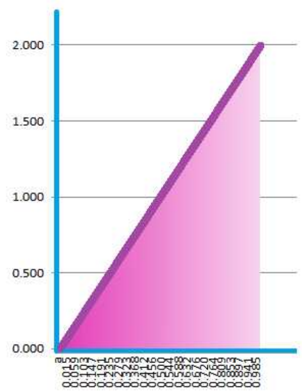
$$f(x) \le \frac{A(x+h)-A(x)}{h} \le f(x+h)\,.$$

Example 1.8.7: triangle

Let's confirm the idea with a familiar shape. Consider

$$f(x) = 2x\,.$$

Let's compute the area function.



We have:

$$\begin{aligned} A(x) &= \int_0^x 2x\,dx \\ &= \text{area of triangle with base } [0,x] \\ &= \frac{1}{2}\,\text{width}\cdot\text{height} \\ &= \frac{1}{2}x\cdot 2x \\ &= x^2\,. \end{aligned}$$

What's the relation between $f(x)=2x$ and $A(x)=x^2$? We know the answer:

$$\left(x^2\right)' = 2x\,,$$

so f is an *antiderivative* of f!

This idea applies to all functions.

Theorem 1.8.8: Fundamental Theorem of Calculus I

Given a continuous function f on $[a,b]$, the function defined by

$$\boxed{F(x) = \int_a^x f\,dx}$$

is an antiderivative of f on (a,b); i.e,

$$F' = f\,.$$

Thus, differentiation cancels the effect of the (variable-end) Riemann integral: $F' = f$.

Exercise 1.8.9

Find a formula for the antiderivative with $F(a) = 1$.

Exercise 1.8.10

Find the derivative of $\int_x^b f\, dx$.

But this is only a half!

Next, we see how the Riemann integral cancels – up to a constant – the effect of differentiation.

Theorem 1.8.11: Fundamental Theorem of Calculus II (Newton-Leibniz Formula)

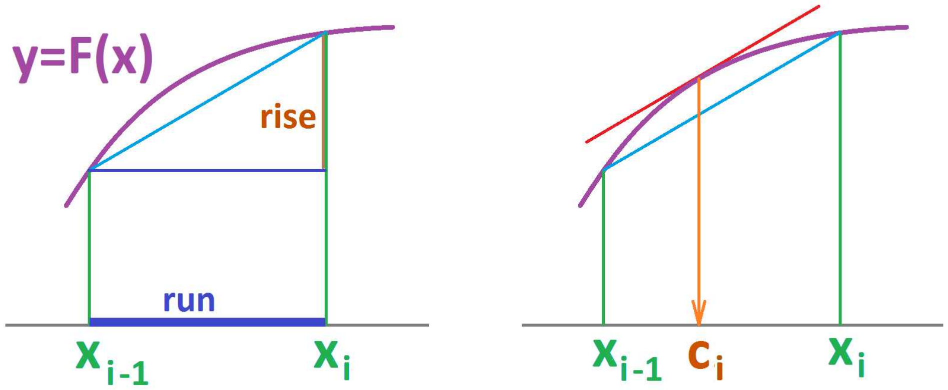
For any integrable function f on $[a, b]$ and any of its antiderivatives F, we have

$$\boxed{\int_a^b f\, dx = F(b) - F(a)}$$

Proof.

We start with a partition P of interval $[a, b]$ with n intervals. The nodes $x_i,\ i = 0, 1, ..., n$, can be arbitrary; we can even choose equal lengths for the intervals: $\Delta x_i = x_i - x_{i-1} = (b - a)/n,\ i = 1, 2, ..., n$. There are no secondary nodes yet; their choice will be dictated by F.

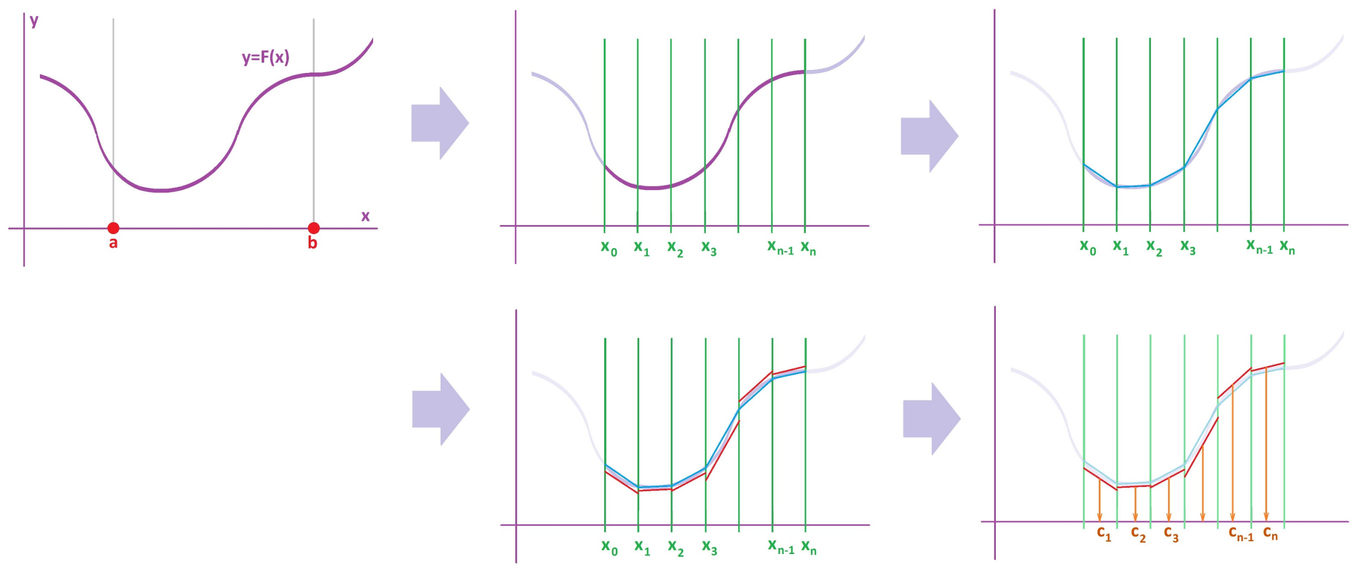
This is how each of the secondary nodes $c_i,\ i = 1, 2, ..., n$, is chosen:



Let's take one interval $[x_{i-1}, x_i]$ and we apply the *Mean Value Theorem* to F: There is a c_i in the interval such that

$$\frac{F(x_i) - F(x_{i-1})}{x_i - x_{i-1}} = F'(c_i)\,.$$

In other words, we find a point within each interval that has the slope of its tangent line equal to the slope of the secant line over the whole interval $[x_{i-1}, x_i]$:



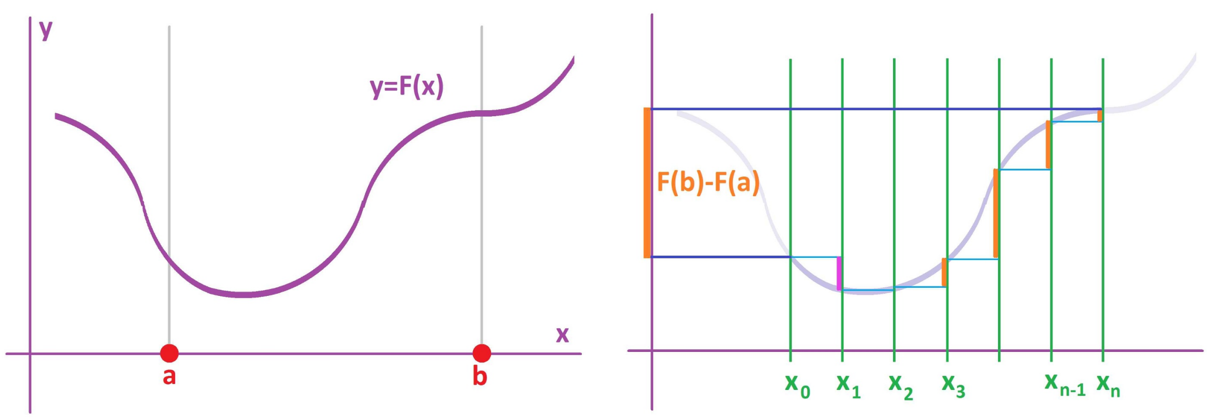
We modify the formula:

$$F(x_i) - F(x_{i-1}) = f(c_i)\,\Delta x\,.$$

Here we have:

- The left-hand side is the *net change* (the rise) of F over the interval $[x_{i-1}, x_i]$.
- The right-hand side is the element of the Riemann sum of f over $[x_{i-1}, x_i]$.

The next step is to express the *total net change* of F over the interval $[a, b]$ as the sum of the net changes over the intervals of the partition:



We now convert those net changes according to the above formula:

$$\begin{array}{llclcl}
F(b) - F(a) & = F(x_n) & = & F(x_n) - F(x_{n-1}) & = & f(c_n)\,\Delta x \\
 & & & +F(x_{n-1}) - F(x_{n-2}) & & +f(c_{n-1})\,\Delta x \\
 & & & +... & & +... \\
 & & & +F(x_{i+1}) - F(x_i) & & +f(c_i)\,\Delta x \\
 & & & +... & & +... \\
 & -F(x_0) & & +F(x_1) - F(x_0) & & +f(c_1)\,\Delta x\,.
\end{array}$$

The last expression is the Riemann sum of f, and since f is integrable, it converges to $\int_a^b f\,dx$ as $\Delta x \to 0$.

Exercise 1.8.12

Use the sigma notation to re-write the last computation.

Exercise 1.8.13

Provide the details for the following derivation of the Fundamental Theorem of Calculus I from the Fundamental Theorem of Calculus II (NLF):

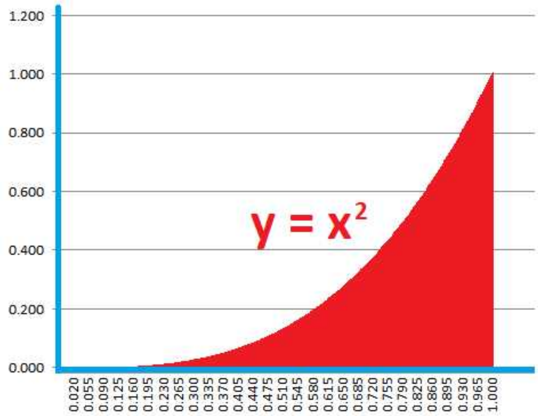
$$\frac{d}{dx}\left(\int_a^x f\,dx\right) = \frac{d}{dx}\left(F(x) - F(a)\right) = F'(x) = f(x)\,.$$

Warning!

Since the theorem applies to *all* antiderivatives, we can omit "$+C$".

Example 1.8.14: parabola

Let's compute the area under $y = x^2$ from 0 to 1:



We did a similar one the hard way, via the Riemann sums, which required a formula for $1+2+...+n$.

Let's test the formula. We need the antiderivative of $f(x) = x^2$. By *recalling* the derivative we computed before:

$$(x^3)' = 3x^2\,.$$

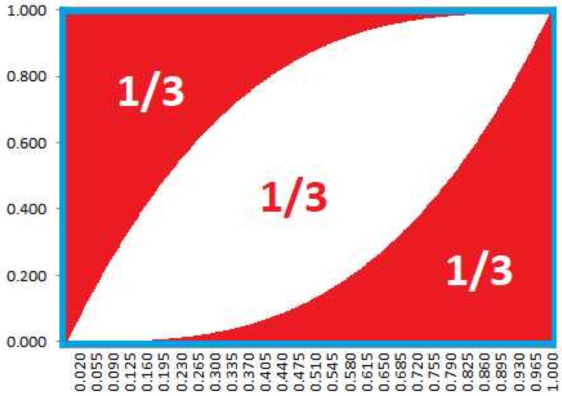
(or using the *Power Formula* from Chapter 2DC-5), we find an antiderivative:

$$F(x) = \frac{x^3}{3}\,.$$

Then, by *NLF*, we have:

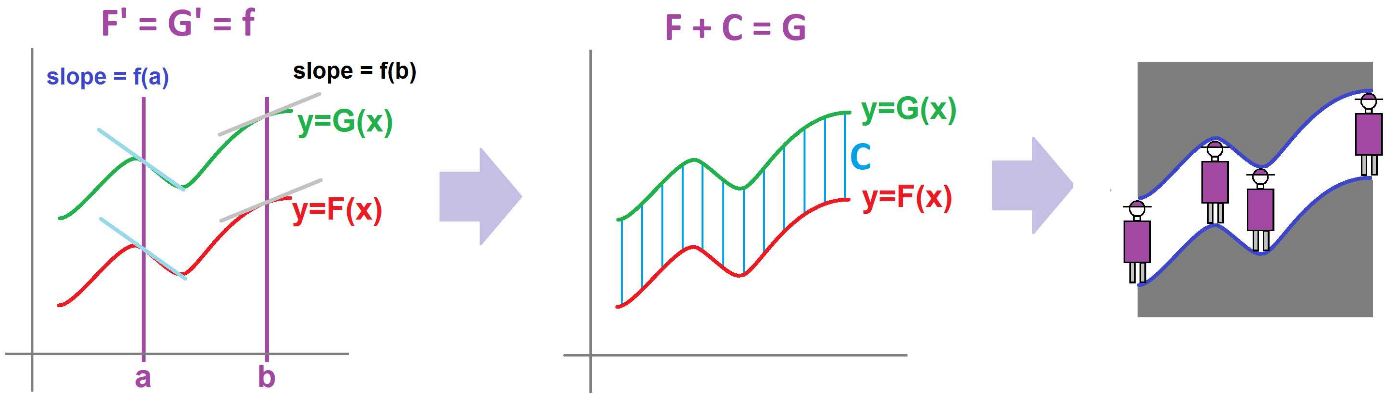
$$\int_0^1 x^2\,dx = F(1) - F(0) = \frac{1^3}{3} - \frac{0^3}{3} = \frac{1}{3}\,.$$

The result is plausible:



The computation has also shown that if we increase the speed from 0 to 1 quadratically over the interval of time $[0, 1]$, then the distance covered is $1/3$.

The Part I supplies us with a special choice of antiderivative – the one that satisfies $F(a) = 0$. The rest, according to Part II, are acquired by vertical shifts: $G = F + C$ (seen in Volume 2, Chapter 2DC-4). The idea is that if the ceiling and the floor of a tunnel are equal at every point, its height is constant:



Recall the *substitution notation*:

$$F(x)\Big|_{x=a} = F(a)\,.$$

In its spirit, we introduce a new notation for the right-hand side of the formula. It is simply *two* substitutions subtracted:

Substitution notation for definite integral

$$\int_a^b f\,dx = F(x)\Big|_a^b = F(x)\Big|_{x=a}^{x=b} = F(b) - F(a)$$

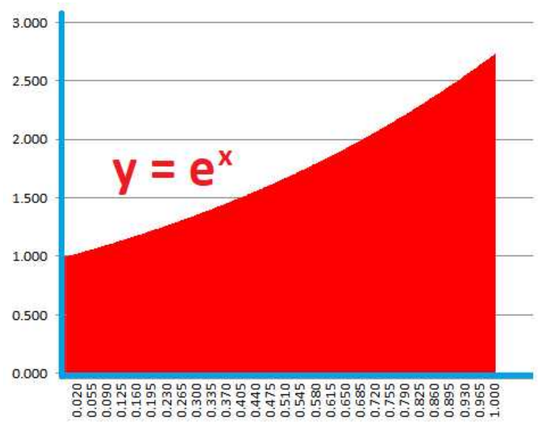
As you can see, the bounds of integration just jump over and are kept for reference. This is how we can record the above computation:

$$\int_0^1 x^2\,dx = \frac{x^3}{3}\Big|_0^1 .$$

Computations of areas become easy – as long as we have the antiderivatives of the function involved.

Example 1.8.15: exponent

Compute the area under $y = e^x$ from 0 to 1:

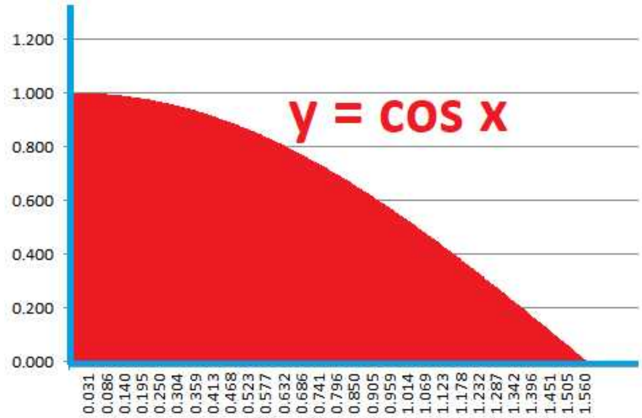


By *NLF*, we have:

$$\int_0^1 e^x\,dx = e^x\Big|_0^1 = e^1 - e^0 = e - 1\,.$$

Example 1.8.16: cosine

Compute the area under $y = \cos x$ from 0 to $\pi/2$:



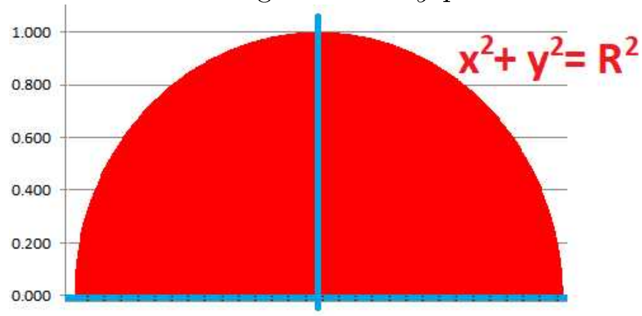
By *NLF*, we have:

$$\int_0^{\frac{\pi}{2}} \cos x \, dx = \sin x \Big|_0^{\frac{\pi}{2}} = \sin \frac{\pi}{2} - \sin 0 = 1 - 0 = 1 \, .$$

Example 1.8.17: circle

What is the area of a circle of radius R? Even though everyone knows the answer, this is the time to *prove* the formula. We also gave an approximate answer at the beginning of this chapter.

We first put the center of the circle at the origin of the xy-plane:



In order to find its area, we represent the *upper half* of the circle as the area under the graph of

$$f(x) = \sqrt{R^2 - x^2}, \ -R \le x \le R \, .$$

Then,

$$\frac{1}{2}\text{Area} \ = \int_{-R}^{R} \sqrt{R^2 - x^2} \, dx \, .$$

What's the antiderivative of this function? Unfortunately, we don't know; it's not on our – very short – list. When this happens, we always do the same thing: We go to the *back of the book* to find a longer list. This *table* contains the relevant formula:

$$\int \sqrt{a^2 - u^2} \, dx = \frac{u}{2}\sqrt{a^2 - u^2} + \frac{a^2}{2} \sin^{-1} \frac{u}{a} + C \, .$$

Now, even though we get the integration formula from elsewhere, once we have it, we can certainly *prove* it – by differentiation. We know these two derivatives from Chapter 2DC-4:

$$\left(\sqrt{a^2 - u^2}\right)' = -\frac{u}{\sqrt{a^2 - u^2}} \quad \text{and} \quad \left(\sin^{-1} y\right)' = \frac{1}{\sqrt{1 - y^2}} \, .$$

Therefore, we have:

$$\begin{aligned}\left(\frac{u}{2}\sqrt{a^2-u^2}+\frac{a^2}{2}\sin^{-1}\frac{u}{a}\right)' &= \frac{1}{2}\sqrt{a^2-u^2}-\frac{u}{2}\frac{u}{\sqrt{a^2-u^2}}+\frac{a^2}{2}\frac{1}{\sqrt{a^2-u^2}}\\ &= \frac{1}{2}\sqrt{a^2-u^2}+\frac{1}{2}\frac{a^2-u^2}{\sqrt{a^2-u^2}}\\ &= \sqrt{a^2-u^2}\,.\end{aligned}$$

Confirmed!

We replace a with R and u with x in the above formula and apply the *Newton-Leibniz formula*:

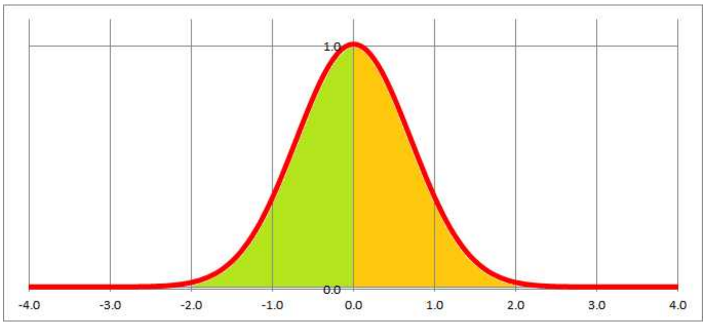
$$\begin{aligned}\frac{1}{2}\text{Area} &= \int_{-R}^{R}\sqrt{R^2-x^2}\,dx\\ &= \frac{x}{2}\sqrt{R^2-x^2}+\frac{R^2}{2}\sin^{-1}\frac{x}{R}\Bigg|_{-R}^{R}\\ &= \left(\frac{R}{2}\sqrt{R^2-R^2}+\frac{R^2}{2}\sin^{-1}\frac{R}{R}\right) \qquad -\left(\frac{-R}{2}\sqrt{R^2-R^2}+\frac{R^2}{2}\sin^{-1}\frac{-R}{R}\right)\\ &= \frac{R^2}{2}\left(\sin^{-1}(1)+0-0-\sin^{-1}(-1)\right)\\ &= \frac{R^2}{2}\left(\pi/2-(-\pi/2)\right)\\ &= \pi\frac{R^2}{2}\,.\end{aligned}$$

We have confirmed the formula! Conversely, we can use the Riemann sum approximation of the area of the unit circle as a way to find approximations of the value of π.

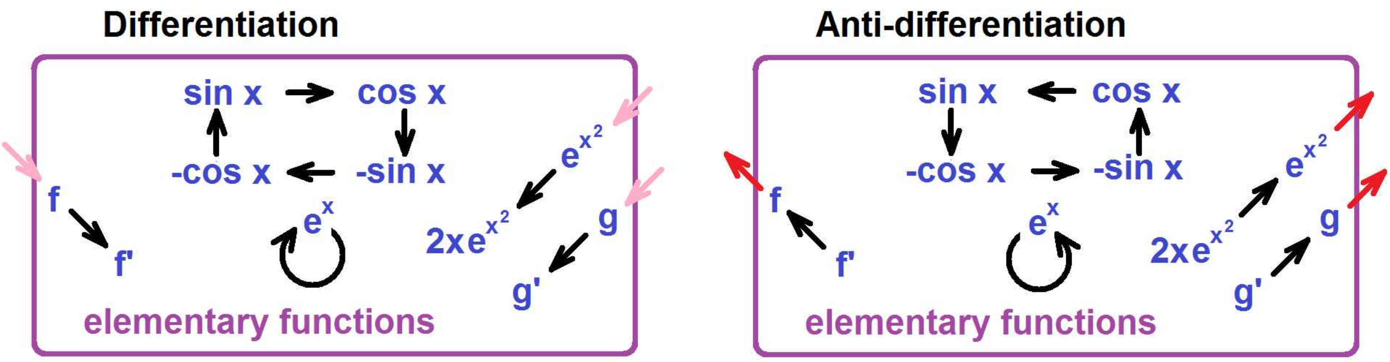
What if we can't find an integration formula for the function in question? There are always larger lists, but it's possible that no book contains the function you need. In fact, we may have to introduce *new functions* such as in the case of the Gauss error function defined as the integral:

$$\operatorname{erf}(x)=\frac{2}{\sqrt{\pi}}\int e^{-x^2}\,dx\,.$$

This function is important as it represents a distribution of probabilities of some common quantities (heights, weights, IQs, etc.) around the average:



Even if we keep adding these new functions to the list of "familiar" functions, there will always remain integrals not on the list. In contrast to differentiation, anti-differentiation will often take us outside of the realm of elementary functions (right):



1.9. *Basic integration*

Recall the procedure for computing the derivative:

$$\begin{tikzcd}
\text{sequence} \arrow[d, "\text{DQ}"] & \text{function on } [a,b] \arrow[l, "\text{sampling}"'] \arrow[d, "\text{derivative}"] \\
\text{sequence} \arrow[r, "\Delta x \to 0"] & \text{function on } [a,b]
\end{tikzcd}$$

The one for the integral has a very similar representation:

$$\begin{tikzcd}
\text{sequence} \arrow[d, "\text{RS}"] & \text{function on } [a,b] \arrow[l, "\text{sampling}"'] \arrow[d, "\text{integral}"] \\
\text{sequence} \arrow[r, "\Delta x \to 0"] & \text{function on } [a,b]
\end{tikzcd}$$

Because of the recursive nature of the Riemann sums, however, the computations are by far more complex or even impossible.

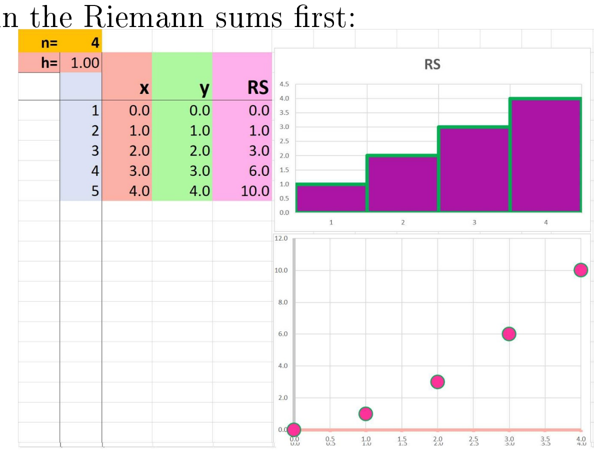
The Fundamental Theorem will give the answer if we can execute the anti-differentiation required.

Example 1.9.1: x

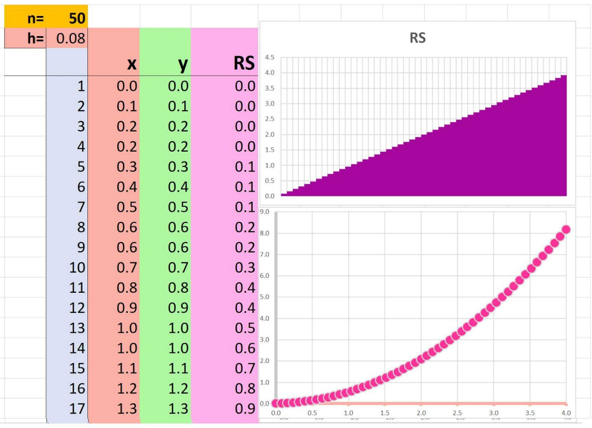
Find the integral of

$$f(x) = x\,.$$

We run the Riemann sums first:



n=	4			
h=	1.00			
	x	y	RS	
1	0.0	0.0	0.0	
2	1.0	1.0	1.0	
3	2.0	2.0	3.0	
4	3.0	3.0	6.0	
5	4.0	4.0	10.0	



n=	50		
h=	0.08		
	x	y	RS
1	0.0	0.0	0.0
2	0.1	0.1	0.0
3	0.2	0.2	0.0
4	0.2	0.2	0.0
5	0.3	0.3	0.1
6	0.4	0.4	0.1
7	0.5	0.5	0.1
8	0.6	0.6	0.2
9	0.6	0.6	0.2
10	0.7	0.7	0.3
11	0.8	0.8	0.4
12	0.9	0.9	0.4
13	1.0	1.0	0.5
14	1.0	1.0	0.6
15	1.1	1.1	0.7
16	1.2	1.2	0.8
17	1.3	1.3	0.9

The answer appear to be quadratic. But

$$(ax^2 + bx + c)' = 2ax + b\,.$$

It works if we choose: $a = 1/2$ and $b = 0$. We conclude:

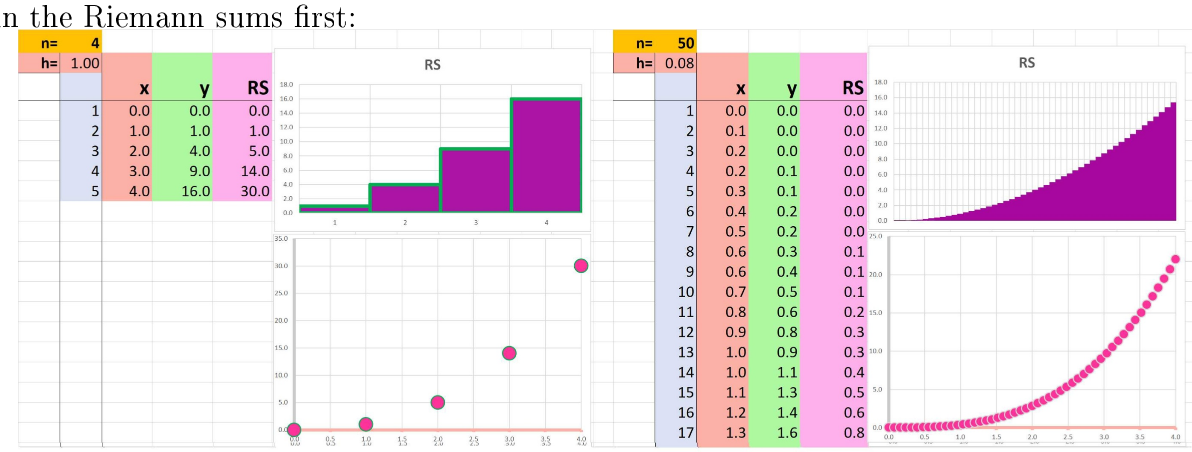
$$\int x\,dx = \frac{1}{2}x^2 + C\,.$$

Example 1.9.2: x^2

Find the integral of

$$f(x) = x^2\,.$$

We run the Riemann sums first:



n=	4			
h=	1.00			
		x	y	RS
	1	0.0	0.0	0.0
	2	1.0	1.0	1.0
	3	2.0	4.0	5.0
	4	3.0	9.0	14.0
	5	4.0	16.0	30.0

n=	50			
h=	0.08			
		x	y	RS
	1	0.0	0.0	0.0
	2	0.1	0.0	0.0
	3	0.2	0.0	0.0
	4	0.2	0.1	0.0
	5	0.3	0.1	0.0
	6	0.4	0.2	0.0
	7	0.5	0.2	0.0
	8	0.6	0.3	0.1
	9	0.6	0.4	0.1
	10	0.7	0.5	0.1
	11	0.8	0.6	0.2
	12	0.9	0.8	0.3
	13	1.0	0.9	0.3
	14	1.0	1.1	0.4
	15	1.1	1.3	0.5
	16	1.2	1.4	0.6
	17	1.3	1.6	0.8

Might it be cubic? Let's try:

$$(ax^3 + bx^2 + cx + d)' = 3ax^2 + 2bx + c\,.$$

It works if we choose: $a = 1/3$, $b = 0$, and $c = 0$. We conclude:

$$\int x^2\,dx = \frac{1}{3}x^3 + C\,.$$

Of course, we can just "reverse" the Power Formula:

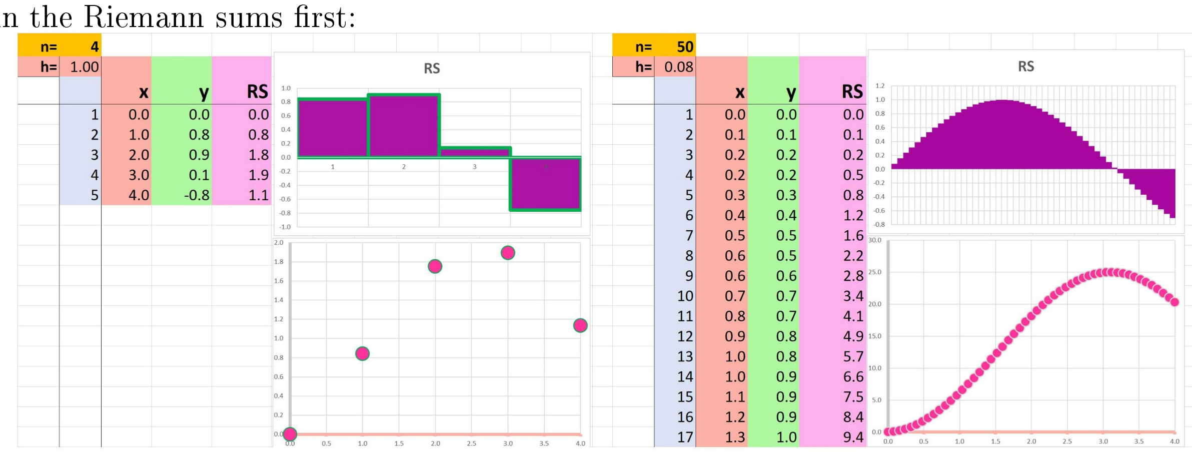
$$\boxed{\int x^n\,dx = \frac{1}{n+1}x^{n+1} + C}$$

Example 1.9.3: $\sin x$ and $\cos x$

Find the integral of

$$f(x) = \sin x\,.$$

We run the Riemann sums first:



n=	4			
h=	1.00			
		x	y	RS
	1	0.0	0.0	0.0
	2	1.0	0.8	0.8
	3	2.0	0.9	1.8
	4	3.0	0.1	1.9
	5	4.0	-0.8	1.1

n=	50			
h=	0.08			
		x	y	RS
	1	0.0	0.0	0.0
	2	0.1	0.1	0.1
	3	0.2	0.2	0.2
	4	0.2	0.2	0.5
	5	0.3	0.3	0.8
	6	0.4	0.4	1.2
	7	0.5	0.5	1.6
	8	0.6	0.5	2.2
	9	0.6	0.6	2.8
	10	0.7	0.7	3.4
	11	0.8	0.7	4.1
	12	0.9	0.8	4.9
	13	1.0	0.8	5.7
	14	1.0	0.9	6.6
	15	1.1	0.9	7.5
	16	1.2	0.9	8.4
	17	1.3	1.0	9.4

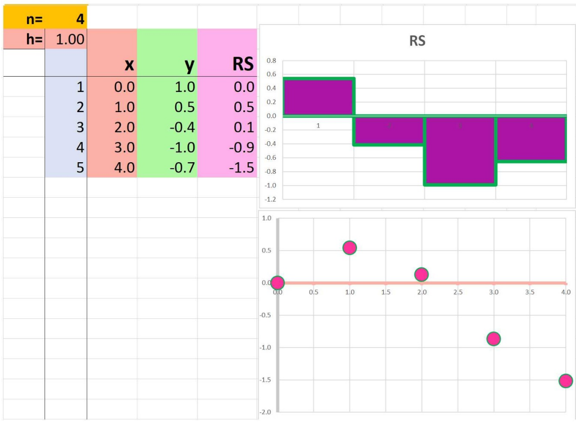
It looks like $\cos x + 1$. Indeed:

$$(\cos x)' = \sin x\,.$$

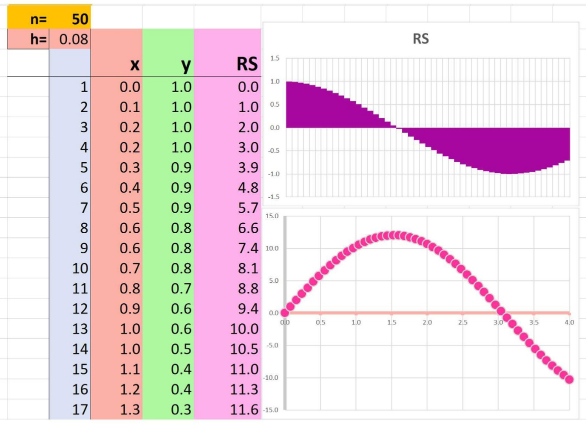
Find the integral of

$$f(x) = \cos x\,.$$

We run the Riemann sums first:



	x	y	RS
1	0.0	1.0	0.0
2	1.0	0.5	0.5
3	2.0	-0.4	0.1
4	3.0	-1.0	-0.9
5	4.0	-0.7	-1.5



	x	y	RS
1	0.0	1.0	0.0
2	0.1	1.0	1.0
3	0.2	1.0	2.0
4	0.2	1.0	3.0
5	0.3	0.9	3.9
6	0.4	0.9	4.8
7	0.5	0.9	5.7
8	0.6	0.8	6.6
9	0.6	0.8	7.4
10	0.7	0.8	8.1
11	0.8	0.7	8.8
12	0.9	0.6	9.4
13	1.0	0.6	10.0
14	1.0	0.5	10.5
15	1.1	0.4	11.0
16	1.2	0.4	11.3
17	1.3	0.3	11.6

It looks like the negative sine. Indeed:

$$(\sin x)' = \cos x\,.$$

Of course, we can just "reverse" a couple of familiar formulas:

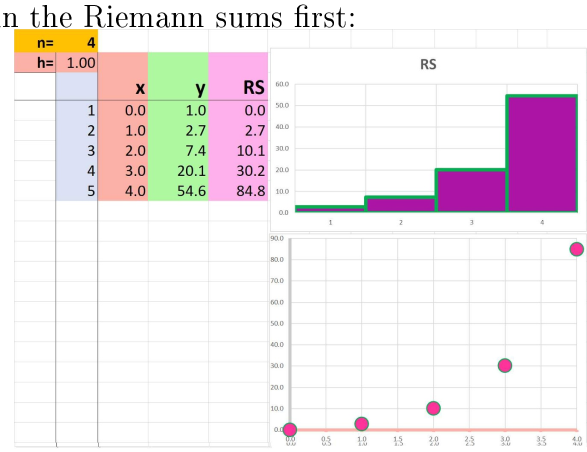
$$\int \sin x\, dx = \cos x + C \qquad \int \cos x\, dx = -\sin x + C$$

Example 1.9.4: e^x

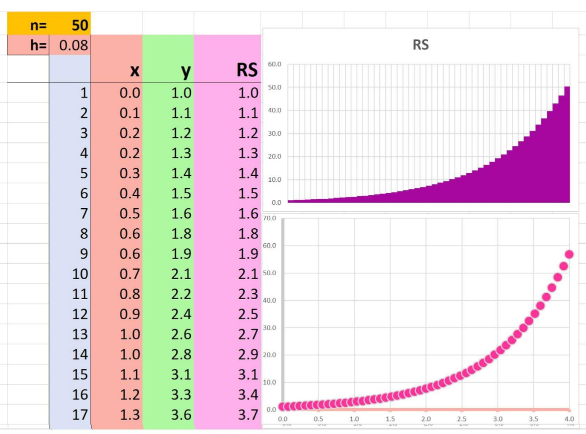
Find the integral of

$$f(x) = e^x\,.$$

We run the Riemann sums first:



	x	y	RS
1	0.0	1.0	0.0
2	1.0	2.7	2.7
3	2.0	7.4	10.1
4	3.0	20.1	30.2
5	4.0	54.6	84.8



	x	y	RS
1	0.0	1.0	1.0
2	0.1	1.1	1.1
3	0.2	1.2	1.2
4	0.2	1.3	1.3
5	0.3	1.4	1.4
6	0.4	1.5	1.5
7	0.5	1.6	1.6
8	0.6	1.8	1.8
9	0.6	1.9	1.9
10	0.7	2.1	2.1
11	0.8	2.2	2.3
12	0.9	2.4	2.5
13	1.0	2.6	2.7
14	1.0	2.8	2.9
15	1.1	3.1	3.1
16	1.2	3.3	3.4
17	1.3	3.6	3.7

It also looks exponential. We conclude:

$$\int e^x\, dx = e^x + C\,.$$

We "reverse" a familiar formula for differentiation:

$$\int e^x\, dx = e^x + C$$

Let's review the main ideas of anti-differentiation.

Suppose we have a function $y = g(x)$ defined at the secondary nodes, c, of a partition. How do we find a function $y = f(x)$ defined at the nodes, x, of the partition so that g is its *difference*:

$$\Delta f(c) = g(c)?$$

In other words, we face an equation:

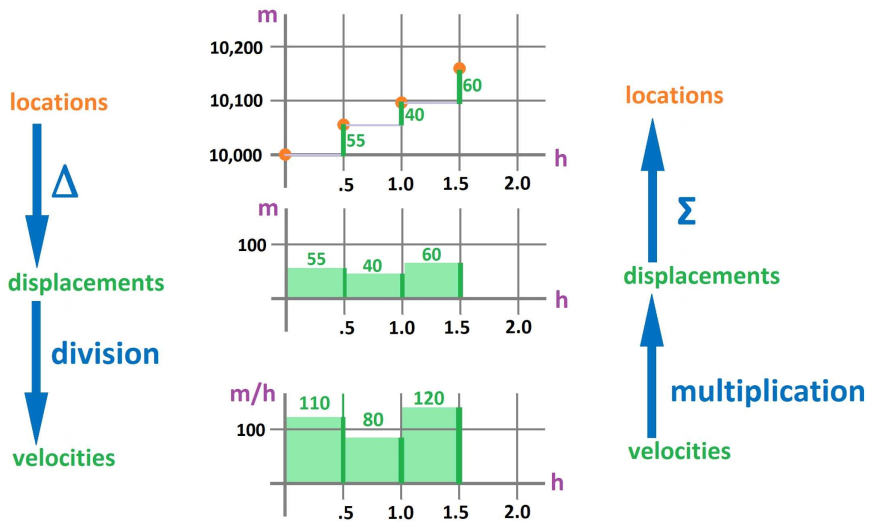
$$\boxed{\text{Solve for } f: \ \Delta f = g}$$

Solving this equation isn't hard. Suppose this function g is known but f isn't, except for one (initial) value: $y_0 = f(a) = f(x_0)$. Then we have:

$$\boxed{f(x_{k+1}) = f(x_k) + g(c_k)}$$

This formula is *recursive*: We need to know the last value of f in order to find the next.

Now, the *difference quotient*. In comparison, there is another operation – division (by Δx) – following the subtraction:



Suppose we have a function $y = v(x)$ (which could be the velocity) defined at the secondary nodes, c, of a partition. How do we find a function $y = p(x)$ (which could be the position) defined at the nodes, x, of the partition so that v is its difference quotient:

$$\frac{\Delta p}{\Delta x}(c) = v(c)?$$

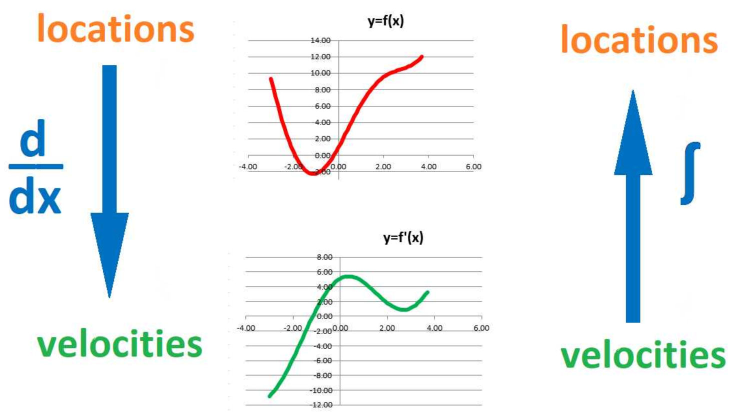
In other words, this is the equation we face:

$$\boxed{\text{Solve for } p: \ \frac{\Delta p}{\Delta x} = v}$$

We just follow exactly the process above. Suppose this function v is known but p isn't, except for one (initial) value: $y_0 = p(a) = p(x_0)$. Then we have:

$$\boxed{p(x_{k+1}) = p(x_k) + v(c_k)\Delta x_k}$$

Now, the *derivative*. We need to find the function from its derivative:



Suppose we have a function $y = v(x)$ (which could be the velocity) defined on an open interval. How do we find a function $y = p(x)$ (which could be the position) defined on this interval so that v is its derivative for each x:

$$\frac{dp}{dx}(x) = v(x)?$$

In other words, this is the equation we face:

$$\boxed{\text{Solve for } p : \ \frac{dp}{dx} = v}$$

This equation isn't as easy to solve as the last. The solution for the difference quotient was built by linking n pieces together. Because the derivative is a limit, it's as if we were to link together infinitely many infinitely small pieces! There is no simple, even recursive, formula. For example, the solution to the following equation cannot be "made" from the functions we are familiar with:

$$\frac{dp}{dx} = e^{x^2}.$$

These three problems are similar to that of finding the *inverse* of a function. This is how inverse functions appear in algebra; they come from solving equations, for x:

$$\begin{array}{llll} x^2 & = 4 \implies & x = 2 & \text{via } \sqrt{\ } \\ 2^x & = 8 \implies & x = 3 & \text{via } \log_2(\ ) \\ \sin x & = 0 \implies & x = 0 & \text{via } \sin^{-1}(\ ) \end{array}$$

In other words, what do we do if we know the output of a function and want to know the input? Initially, we can only *recollect* a past experience with a function! For a repeated use, we develop the *inverse* of the function.

Similarly, what do we do if we know the result of differentiation and want to know where it came from? Initially, we can only *recollect* a past experience with differentiation:

$$\begin{array}{llll} f' & = 2x & \implies f & = x^2 \\ f' & = \cos x & \implies f & = \sin x \\ f' & = e^x & \implies f & = e^x \end{array}$$

For a repeated use, we will need to develop the *inverse* of the function.

We illustrate the idea of anti-differentiation with a diagram:

$$\begin{array}{ccccc} x^2 & \to & \boxed{\dfrac{d}{dx}} & \to & 2x \\ 2x & \to & \boxed{\left(\dfrac{d}{dx}\right)^{-1}} & \to & x^2 \end{array}$$

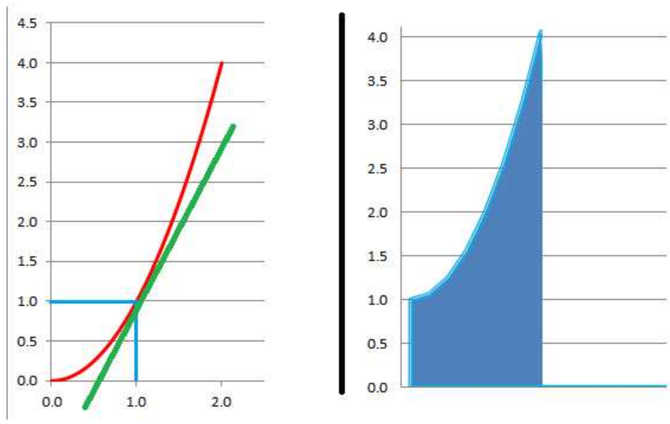
Of course, there are more solutions:

$$\begin{array}{ccl} & & x^2+1 \\ & \nearrow & \dots \\ 2x & \to & x^2 \\ & \searrow & \dots \\ & & x^2-1 \end{array}$$

As a function, a function of functions, $\frac{d}{dx}$ isn't one-to-one. We use "an" because there are many antiderivatives for each function.

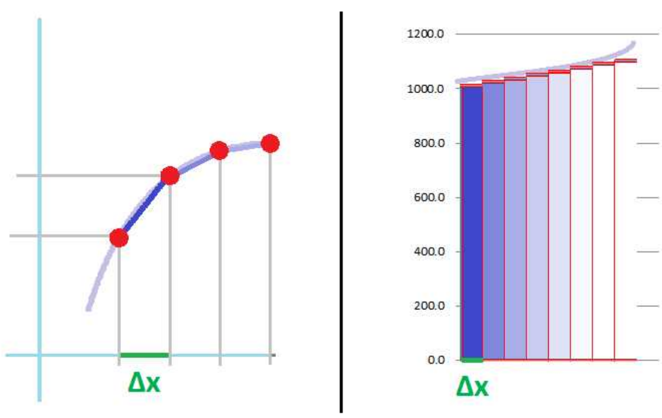
The Fundamental Theorem is the culmination of our study of calculus, so far. The milestones of this study up to this point are outlined below:

The Tangent Problem	The Area Problem



Approximations:

the difference quotient	the Riemann sum
$\dfrac{\Delta F}{\Delta x}$	$\Sigma f \cdot \Delta x$
the slopes of the secant lines	the total area of the rectangles
on the graph of $y = F(x)$	under the graph of $y = f(x)$



With variable locations, the limits of these approximations, as $\Delta x \to 0$, are the following:

the derivative function	the indefinite integral

According to the Fundamental Theorems, the operations of differentiation and integration cancel each other:

$$\text{FTCI:}\quad f \ \to\ \boxed{\int \square\, dx}\ \to\ F\ \to\ \boxed{\frac{d}{dx}\square}\ \to\ f$$

$$\text{FTCII:}\quad F \ \to\ \boxed{\frac{d}{dx}\square}\ \to\ f\ \to\ \boxed{\int \square\, dx}\ \to\ F+C$$

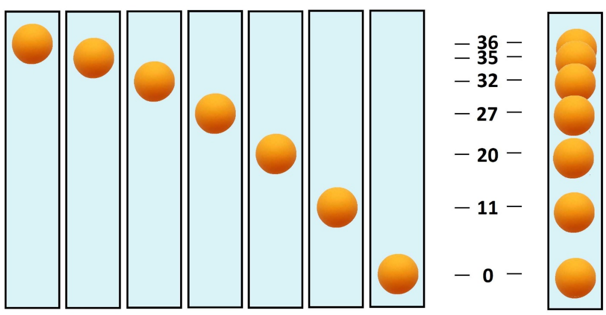
Warning!

Integration is a true function only when an extra condition, such as $F(a) = 0$, is imposed.

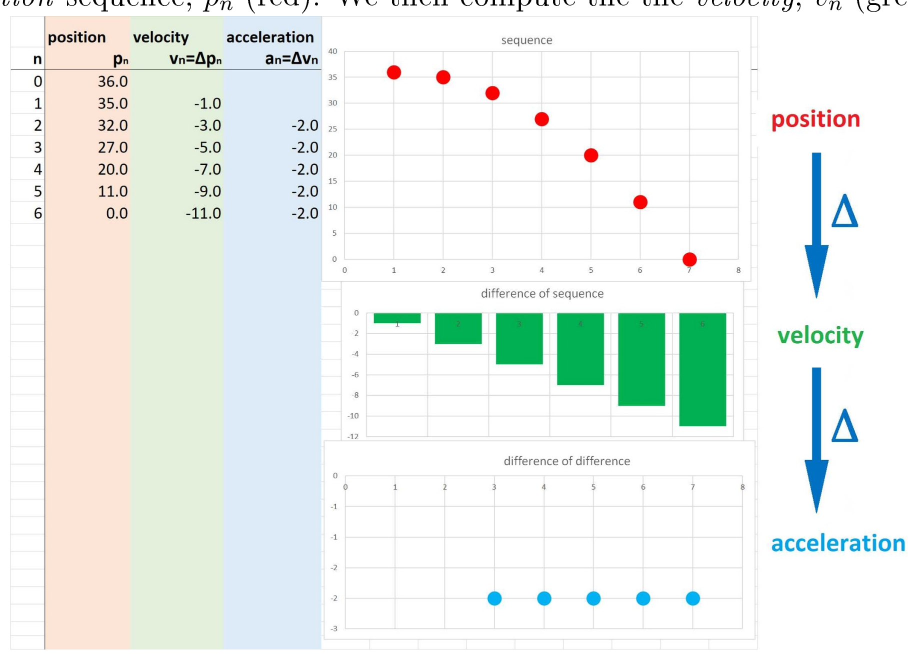
1.10. Free fall

Example 1.10.1: moving ball

Let's review an example from Volume 2. The velocity of a ball thrown up in the air is constantly changed by the gravity. Imagine that we have this experimental data of the heights of a ping-pong ball falling down recorded about every 0.1 second, measured in inches:



We plot the *location* sequence, p_n (red). We then compute the the *velocity*, v_n (green):



n	position p_n	velocity $v_n=\Delta p_n$	acceleration $a_n=\Delta v_n$
0	36.0		
1	35.0	-1.0	
2	32.0	-3.0	-2.0
3	27.0	-5.0	-2.0
4	20.0	-7.0	-2.0
5	11.0	-9.0	-2.0
6	0.0	-11.0	-2.0

We compute the *acceleration*, a_n (blue), too. It appears constant.

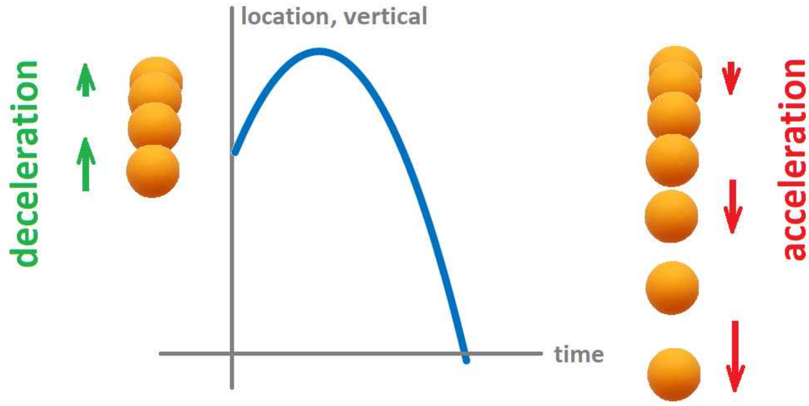
Let's accept the premise we've put forward:

▶ *The acceleration of free fall is constant.*

Then we can try to predict the behavior of an object thrown in the air – from any initial height and with any initial velocity:

▶ We use our knowledge of the acceleration to derive the velocity, and then derive the position of the object in time.

We plot the position against time:



We used these difference quotient formulas to find the velocity from the position and then the acceleration from the velocity:

$$p_n \quad \xrightarrow{\text{DQ}} \quad v_n = \frac{\Delta p}{\Delta t} = \frac{p_{n+1} - p_n}{h} \quad \xrightarrow{\text{DQ}} \quad a_n = \frac{\Delta v}{\Delta t} = \frac{v_{n+1} - v_n}{h}$$

Here h is the increment of time. The dependence of the velocity on the position and of the acceleration on the velocity is, of course, identical.

To create formulas for a simulation of free fall, the derivation goes in the opposite direction:

- the velocity from the acceleration, and then
- the location from the acceleration.

The two formulas above are solved as equations for p_{n+1} and v_{n+1} respectively:

$$\begin{aligned} v_n = \frac{p_{n+1} - p_n}{h} \quad &\Longrightarrow \quad p_{n+1} = p_n + hv_n \\ a_n = \frac{v_{n+1} - v_n}{h} \quad &\Longrightarrow \quad v_{n+1} = v_n + ha_n \end{aligned}$$

The dependence of the velocity on the acceleration and of the position on the velocity is, of course, identical.

Warning!

Unlike the former, these are *recursive* sequences.

Example 1.10.2: free fall

▶ *Problem:* From a 100-foot building, a ball is thrown up at 50 feet per second in such a way that it falls on the ground. How high will the ball go?

We use the same spreadsheet formula for the velocity and position:

```
=R[-1]C+R[-1]C[-1]*R2C1
```

Now in the specific case of *free fall*, there is just one force, the gravity, and the vertical acceleration is known (from a physics textbook) to be $a = -g$, where g is the gravitational constant:

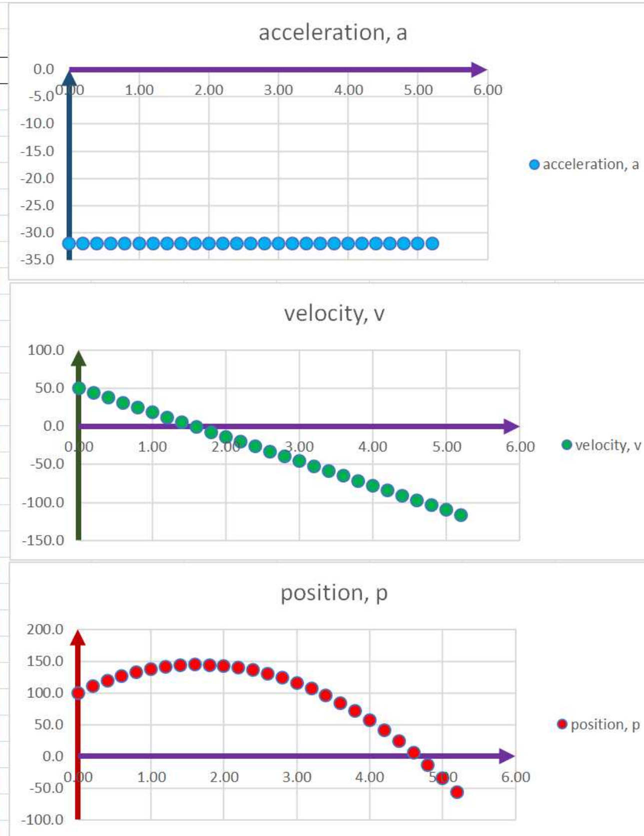
$$g = 32 \text{ ft/sec}^2 .$$

Next, we acquire the initial conditions:

- The initial location is given by: $p_0 = 100$.
- The initial velocity is given by: $v_0 = 50$.

We use the formulas to evaluate the location every $h = 0.20$ second. This is what the graphs look like:

h				
0.20	t	a	v	p
intial:	*0.00*	*-32.00*	*50.00*	*100.00*
	0.20	-32.00	43.60	110.00
	0.40	-32.00	37.20	118.72
	0.60	-32.00	30.80	126.16
	0.80	-32.00	24.40	132.32
	1.00	-32.00	18.00	137.20
	1.20	-32.00	11.60	140.80
	1.40	-32.00	5.20	143.12
	1.60	-32.00	-1.20	144.16
	1.80	-32.00	-7.60	143.92
	2.00	-32.00	-14.00	142.40
	2.20	-32.00	-20.40	139.60
	2.40	-32.00	-26.80	135.52
	2.60	-32.00	-33.20	130.16
	2.80	-32.00	-39.60	123.52
	3.00	-32.00	-46.00	115.60
	3.20	-32.00	-52.40	106.40
	3.40	-32.00	-58.80	95.92
	3.60	-32.00	-65.20	84.16
	3.80	-32.00	-71.60	71.12
	4.00	-32.00	-78.00	56.80
	4.20	-32.00	-84.40	41.20
	4.40	-32.00	-90.80	24.32
	4.60	-32.00	-97.20	6.16
	4.80	-32.00	-103.60	-13.28
	5.00	-32.00	-110.00	-34.00
	5.20	-32.00	-116.40	-56.00
	5.40	-32.00	-122.80	-79.28
	5.60	-32.00	-129.20	-103.84
	5.80	-32.00	-135.60	-129.68



By simply examining the data, we can solve various problems about this experiment:

1. To find the highest elevation, we look at the row with p. The largest value seems to be close to $y = 144$ feet.
2. To find when the ball hits the ground, we scroll down to find the row with p close to 0. It happens sometime close to $t = 4.7$ seconds.
3. To find how fast the ball hits the ground, we scroll down again to find the row with p close to 0 and look up the value of v. It is close to $v = 200$ feet per second.

Of course, we can decrease the time increment $h = \Delta x$ and get more accurate answers.

With our free fall spreadsheet, we can ask and answer a variety of other questions about such motion (how hard it hits the ground, etc.). However, we can only do one example at a time! The conclusions we draw are specific to these initial conditions. They are also specific to the magnitude of the gravity, and would be different on Mars. And so on.

The results are also dependent on *our* choice of the increment $h = \Delta x$. This is why we now proceed to the *continuous* case and take the limit:

$$h = \Delta x \to 0\,.$$

This time, instead of sequences, we have these *functions of time*:

1. p is the height, the vertical location.
2. v is the vertical velocity.
3. a is the vertical acceleration.

We have first:

$$v = p'\,,\ a = v'\,,$$

and, accordingly,

$$p = \int v\,dx\,,\ v = \int a\,dx\,.$$

Now the specific case of free fall:

$$a = -g\,.$$

We know that:

1. The derivative of a quadratic polynomial is linear.
2. The derivative of a linear polynomial is constant.

And, conversely:

1. The only function the derivative of which is linear is a quadratic polynomial.
2. The only function the derivative of which is constant is a linear polynomial.

From the latter two, we derive:

$$a = a(t) \text{ is constant } \implies v = v(t) \text{ is linear} \implies p = p(t) \text{ is quadratic.}$$

More precisely, we have:

$$\begin{array}{llll} 1. & p(t) & & = At^2 + Bt + C \\ 2. & v(t) & = p'(t) & = 2At + B \\ 3. & a(t) & = v'(t) & = 2A \end{array}$$

What makes these coefficients, A, B, and C, specific are the *initial conditions* and the acceleration:

1. p_0 is the initial height, $p_0 = p(0)$.
2. v_0 is the initial vertical component of velocity, $v_0 = v(0)$.
3. $-g$ is the acceleration, $-g = a(t)$.

We use the third item:

$$2A = -g \implies A = -\frac{g}{2}\,.$$

We use the second item:

$$2At + B\Big|_{t=0} = B = v_0\,.$$

We use the first item:

$$At^2 + Bt + C\Big|_{t=0} = C = p_0\,.$$

Then, our model of motion takes its final form:

$$\boxed{\begin{array}{lllll} 1. & p(t) & = p_0 & +v_0 t & -\frac{1}{2}gt^2 \\ 2. & v(t) & = & v_0 & -gt \\ 3. & a(t) & = & & -g \end{array}}$$

The problem is solved with absolute accuracy!

Example 1.10.3: free fall

► *Problem:* From a 100-foot building, a ball is thrown up at 50 feet per second in such a way that it falls on the ground. How high will the ball go?

We have:

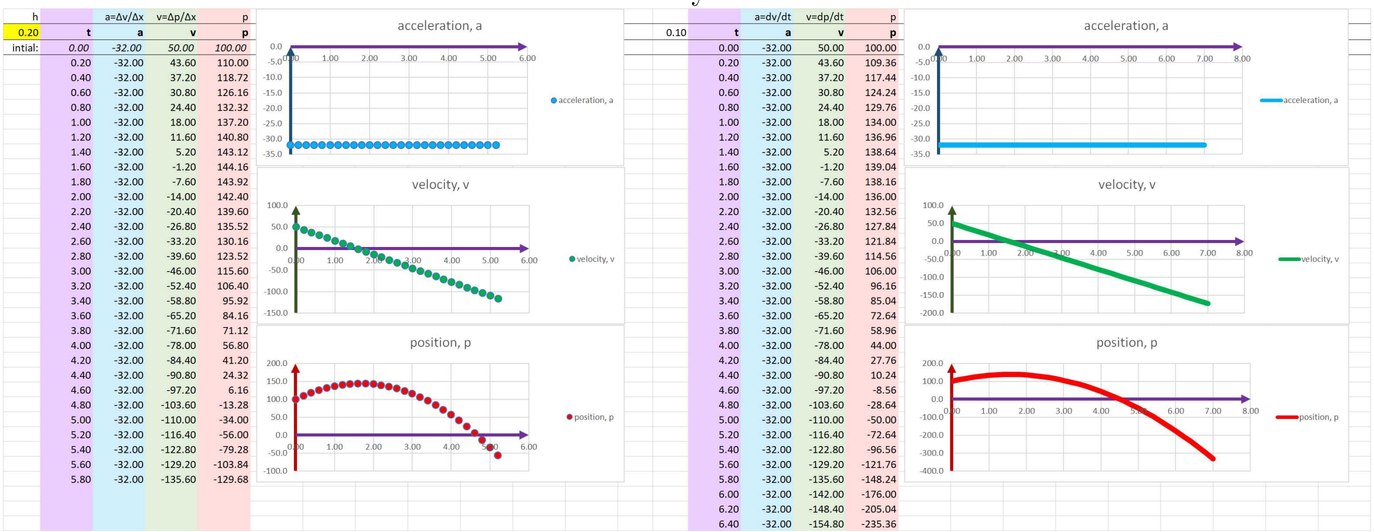
$$p_0 = 100, \ v_0 = 50 \, .$$

Our equation becomes:

$$p \quad = 100 \quad +50t \quad -16t^2 \, .$$

In contrast to the discrete case, the formula for the position isn't recursive but direct and explicit!

Before we utilize the explicit algebraic representation, we visualize the results by plotting the graph of this function next to the one obtained recursively:



h		a=Δv/Δx	v=Δp/Δx	p
0.20	**t**	**a**	**v**	**p**
intial:	*0.00*	*-32.00*	*50.00*	*100.00*
	0.20	-32.00	43.60	110.00
	0.40	-32.00	37.20	118.72
	0.60	-32.00	30.80	126.16
	0.80	-32.00	24.40	132.32
	1.00	-32.00	18.00	137.20
	1.20	-32.00	11.60	140.80
	1.40	-32.00	5.20	143.12
	1.60	-32.00	-1.20	144.16
	1.80	-32.00	-7.60	143.92
	2.00	-32.00	-14.00	142.40
	2.20	-32.00	-20.40	139.60
	2.40	-32.00	-26.80	135.52
	2.60	-32.00	-33.20	130.16
	2.80	-32.00	-39.60	123.52
	3.00	-32.00	-46.00	115.60
	3.20	-32.00	-52.40	106.40
	3.40	-32.00	-58.80	95.92
	3.60	-32.00	-65.20	84.16
	3.80	-32.00	-71.60	71.12
	4.00	-32.00	-78.00	56.80
	4.20	-32.00	-84.40	41.20
	4.40	-32.00	-90.80	24.32
	4.60	-32.00	-97.20	6.16
	4.80	-32.00	-103.60	-13.28
	5.00	-32.00	-110.00	-34.00
	5.20	-32.00	-116.40	-56.00
	5.40	-32.00	-122.80	-79.28
	5.60	-32.00	-129.20	-103.84
	5.80	-32.00	-135.60	-129.68

		a=dv/dt	v=dp/dt	p
0.10	**t**	**a**	**v**	**p**
	0.00	-32.00	50.00	100.00
	0.20	-32.00	43.60	109.36
	0.40	-32.00	37.20	117.44
	0.60	-32.00	30.80	124.24
	0.80	-32.00	24.40	129.76
	1.00	-32.00	18.00	134.00
	1.20	-32.00	11.60	136.96
	1.40	-32.00	5.20	138.64
	1.60	-32.00	-1.20	139.04
	1.80	-32.00	-7.60	138.16
	2.00	-32.00	-14.00	136.00
	2.20	-32.00	-20.40	132.56
	2.40	-32.00	-26.80	127.84
	2.60	-32.00	-33.20	121.84
	2.80	-32.00	-39.60	114.56
	3.00	-32.00	-46.00	106.00
	3.20	-32.00	-52.40	96.16
	3.40	-32.00	-58.80	85.04
	3.60	-32.00	-65.20	72.64
	3.80	-32.00	-71.60	58.96
	4.00	-32.00	-78.00	44.00
	4.20	-32.00	-84.40	27.76
	4.40	-32.00	-90.80	10.24
	4.60	-32.00	-97.20	-8.56
	4.80	-32.00	-103.60	-28.64
	5.00	-32.00	-110.00	-50.00
	5.20	-32.00	-116.40	-72.64
	5.40	-32.00	-122.80	-96.56
	5.60	-32.00	-129.20	-121.76
	5.80	-32.00	-135.60	-148.24
	6.00	-32.00	-142.00	-176.00
	6.20	-32.00	-148.40	-205.04
	6.40	-32.00	-154.80	-235.36

The latter is a sampling of the exact solution.

Just as in the last example, we can use this plot to solve a variety of problems about this motion. Let's revisit the two problems about this specific throw we solved numerically. They are solved the same way:

1. To find the highest elevation, we look at the row with p. The largest value seems to be close to $y = 139$ feet.
2. To find when the ball hits the ground, we scroll down to find the row with p close to 0. It happens sometime close to $t = 4.5$ seconds.

These are approximate solutions as they come from a sampled function. The formula, however, gives us a way to answer the questions with absolute accuracy. We can even avoid differentiation.

For the first problem, we realize that $p = -16t^2 + 50t + 100$ is a parabola! And the vertex of $y = ax^2 + bx + c$ is at $x = -b/a$ (Volume 1, Chapter 1PC-4). Therefore, the highest point is reached at time

$$t = -50/(-2 \cdot 16) = 1.5625 \, .$$

Then, the highest elevation is

$$p = 100 + 50 \cdot 1.5625 - 16 \cdot 1.5625^2 = 139.0625 \, .$$

The result matches our estimate!

For the second problem, the altitude at the end is 0, so to find *when* it happened, we set $p = 0$, or

$$-16t^2 + 50t + 100 = 0 \, ,$$

and solve for t. According to the *Quadratic Formula* (seen in Volume 1, Chapter 1PC-4), we have:

$$t = \frac{-50 - \sqrt{50^2 - 4(-16)100}}{2 \cdot (-16)} \approx 4.5106 \, .$$

The result matches our estimate!

Exercise 1.10.4

What happened to $\pm$?

Exercise 1.10.5

How high does the projectile go in the above example?

Exercise 1.10.6

Using the above example, how long will it take for the projectile to reach the ground if fired *down*?

Exercise 1.10.7

Use the above model to determine how long it will take for an object to reach the ground if it is dropped. Make up your own questions about the situation and answer them. Repeat.

With our simple model of motion, all possible scenarios have been found in the form of an explicit formula!

We can't expect to avoid approximations though:

1. There may be no explicit formula for the velocity, and the position, when the formula for the acceleration as a function of time is complex enough. We have to go back to our discrete model.
2. Even when there are explicit formulas for the velocity and the position, the equation may have no explicit formula for the solution (the time in flight) when the former is too complex. We have to seek an approximate solution.
3. Even in our explicit formula for the time in flight in the last example, the square root will still have to be approximated.

Chapter 2: Integration

Contents

2.1. Properties of Riemann integrals

Consider this *obvious* statement about motion:

- ► "If my speed is zero, I am standing still (and vice versa)."

Proving the mathematical version of this statement will confirm that our theory matches the reality and the common sense. We did this in Volume 2. As a review, we will go through these three stages again:

1. the difference
2. the difference quotient
3. the derivative

This time, we will accompany each of those statements with its equivalent in terms of, respectively:

1. the sum
2. the Riemann sum
3. the Riemann integral

Theorem 2.1.1: Constancy vs. Zero Difference

A function defined at the primary nodes of a partition of an interval has a zero difference (for all secondary nodes in the partition) if and only if this function is constant.

In other words, we have:

$$\Delta F = 0 \iff F = \text{ constant}.$$

Corollary 2.1.2: Zero Function vs. Constant Sum

A function defined at the secondary nodes of a partition of an interval has a constant sum (over all primary nodes of the partition) if and only if this function is zero.

In other words, we have:

$$k = 0 \iff \Sigma k = \text{ constant}.$$

We can think of F as the incremental position and k as the displacement.

Exercise 2.1.3

Derive the latter result from the former.

Next, we divide Δx or factor out Δx (i.e., $k = f\Delta x$):

Theorem 2.1.4: Constancy vs. Zero Difference Quotient

A function defined at the primary nodes of a partition of an interval has a zero difference quotient (for all secondary nodes in the partition) if and only if this function is constant.

In other words, we have:

$$\frac{\Delta F}{\Delta x} = 0 \iff F = \text{ constant}\,.$$

Corollary 2.1.5: Zero Function vs. Constant Riemann Sum

A function defined at the secondary nodes of a partition of an interval has a constant Riemann sum (over all primary nodes of the partition) if and only if this function is zero.

In other words, we have:

$$f = 0 \iff \Sigma f \cdot \Delta x = \text{ constant}.$$

We can think of F as the incremental position and f as the velocity.

Exercise 2.1.6

Derive the latter result from the former.

Next, we make $\Delta x \to 0$:

Theorem 2.1.7: Constancy vs. Zero Derivative

A differentiable on an interval function has a zero derivative over the interval if and only if this function is constant.

In other words, we have:

$$\frac{dF}{dx} = 0 \iff F = \text{ constant.}$$

Corollary 2.1.8: Zero Function vs. Constant Riemann Integral

An integrable function defined on an interval has a constant Riemann integral if and only if this function is zero.

In other words, we have:

$$f = 0 \iff \int_a^x f\,dx = \text{ constant.}$$

We can think of F as the continuously changing position and f as the instantaneous velocity.

Exercise 2.1.9

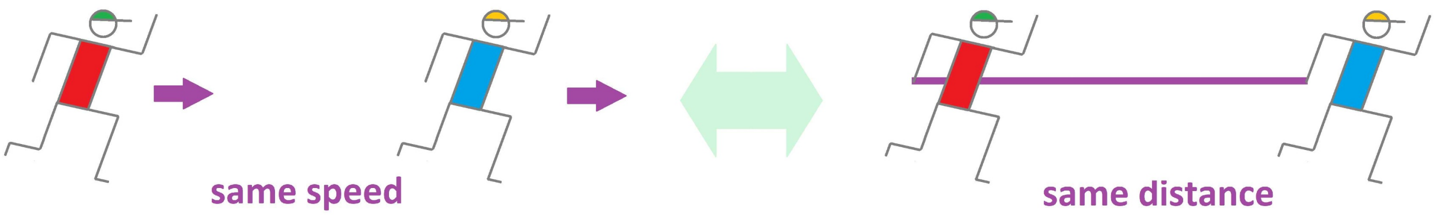
Derive the latter result from the former.

Suppose now that there are *two* runners running with the same speed; what can we say about their mutual locations?

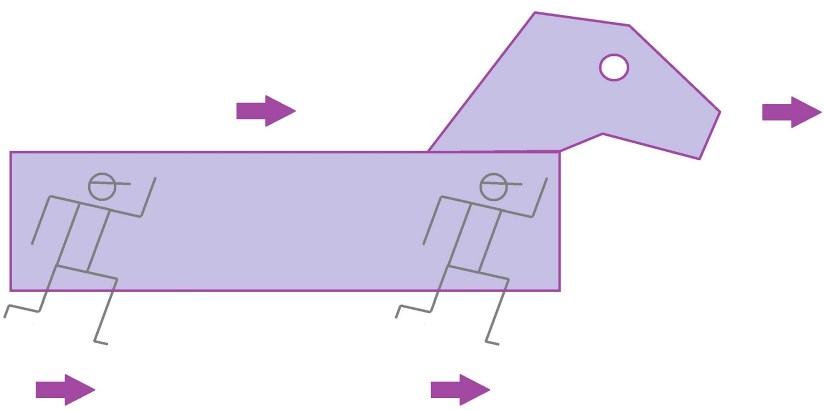
They are not, of course, standing still, but they *are* still relative to each other! We have a slightly less obvious fact about motion:

- ▶ "If two runners run with the same speed, the distance between them doesn't change (and vice versa)".

It's as if they are holding the two ends of a pole without pulling or pushing:



The fact remains valid even if they speed up and slow down all the time. They move as if a single body:



Once again, for functions $y = F(x)$ and $y = G(x)$ representing their positions, we can restate this idea mathematically in order to confirm that our theory makes sense.

We follow the same three stages starting with the differences:

Theorem 2.1.10: Anti-differentiation for Differences

Two functions defined at the primary nodes of a partition of an interval have the same differences (for all secondary nodes of the partition) if and only if the two differ by a constant.

In other words, we have:

$$\Delta F = \Delta G \iff F - G = \text{ constant}.$$

Corollary 2.1.11: Differentiation for Sums

Two functions defined at the secondary nodes of a partition of an interval are equal if and only if the sums of the two differ by a constant (on all primary nodes of the partition).

In other words, we have:

$$f = g \iff \Sigma f - \Sigma g = \text{ constant}.$$

We can think of F and G as the incremental positions and f and g as the displacements.

Exercise 2.1.12

Derive the latter result from the former.

Next, we divide Δx or factor out Δx:

Theorem 2.1.13: Anti-differentiation for Different Quotients

Two functions defined at the nodes of a partition of an interval have the same difference quotient if and only if the two differ by a constant.

In other words, we have:

$$\frac{\Delta F}{\Delta x} = \frac{\Delta G}{\Delta x} \iff F - G = \text{ constant}.$$

Corollary 2.1.14: Differentiation for Riemann Sums

Two functions defined at the secondary nodes of a partition of an interval are equal if and only if the Riemann sums of the two differ by a constant (on all primary nodes of the partition).

In other words, we have:

$$f = g \iff \Sigma f \cdot \Delta x - \Sigma g \cdot \Delta x = \text{ constant}.$$

We can think of F and G as the incremental positions and f and g as the velocities.

Exercise 2.1.15

Derive the latter result from the former.

Next, we make $\Delta x \to 0$:

Theorem 2.1.16: Anti-differentiation for Derivatives

Two functions differentiable on an interval have the same derivative if and only if the two differ by a constant.

In other words, we have:

$$\frac{d}{dx}F = \frac{d}{dx}G \iff F - G = \text{ constant.}$$

Corollary 2.1.17: Differentiation for Integrals

Two functions integrable on an interval have the same Riemann integral if and only if the Riemann integrals of the two differ by a constant.

In other words, we have:

$$f = g \iff \int_a^x f\,dx - \int_a^x g\,dx = \text{ constant.}$$

We can think of F and G as the continuously changing positions and f and g as the instantaneous velocities.

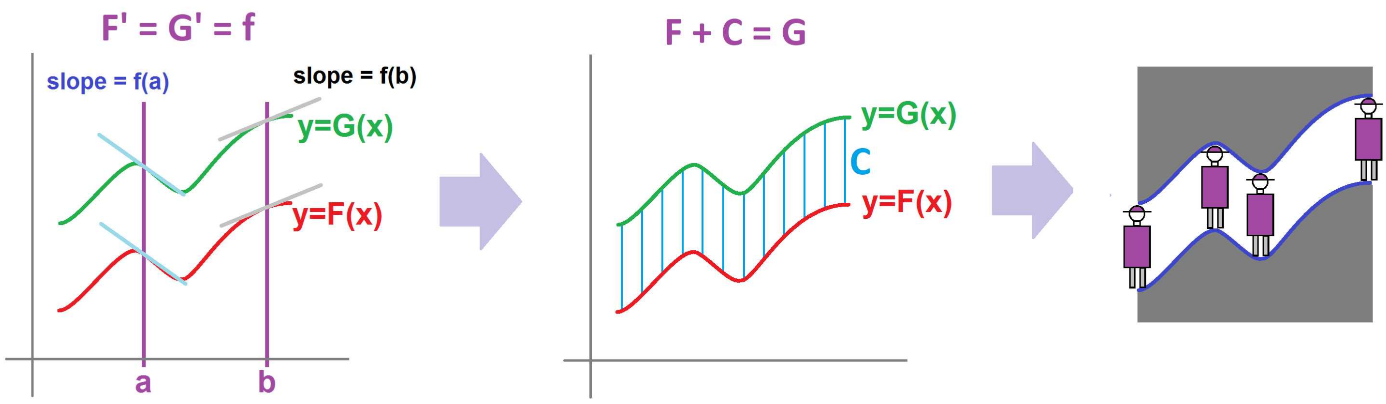
Exercise 2.1.18

Derive the latter result from the former.

In addition to the motion interpretation, there is also one in terms of geometry. The last theorem says:

▶ If the graphs of $y = F(x)$ and $y = G(x)$ have parallel tangent lines for every value of x, then the graph of F is a vertical shift of the graph of G (and vice versa).

We can understand this idea if we imagine a tunnel and a person whose head is touching the ceiling. If the ceiling is sloped down, should he be concerned about hitting his head? Not if the floor is sloped down as much:



In other words, if the slope of the tunnel's top is equal to the slope of the bottom at every location, then the height of the tunnel remains the same throughout its length.

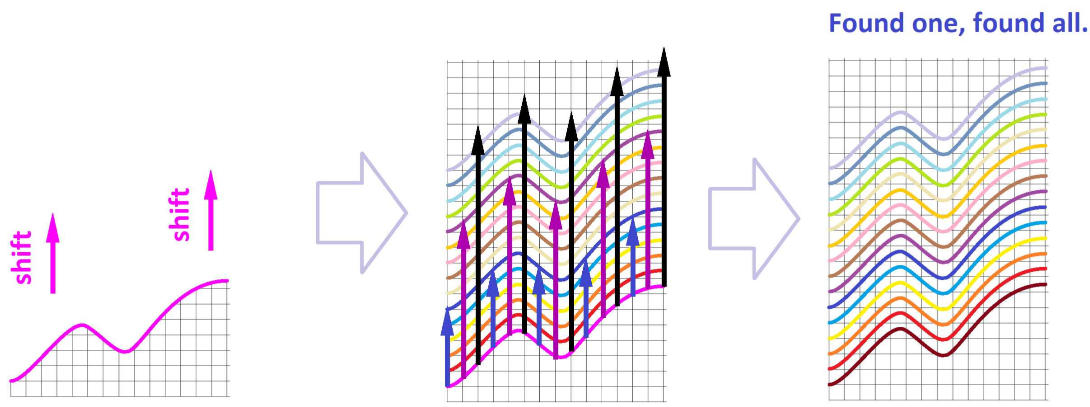
Based on the theorem, we can now update this list of simple but important facts:

info about F		info about F'	info about $\int f\,dx$		info about f
F is constant.	$\iff$	F' is zero.	$\int f\,dx$ is constant.	$\iff$	f is zero.
F is linear.	$\iff$	F' is constant.	$\int f\,dx$ is linear.	$\iff$	f is constant.
F is quadratic.	$\iff$	F' is linear.	$\int f\,dx$ is quadratic.	$\iff$	f is linear.

We use the last two facts to justify our analysis of *free flight*:

Functions of time	
The acceleration is constant.	$\Longrightarrow$
The velocity is linear.	$\Longrightarrow$
The location is quadratic.	

Another important conclusion is that there are infinitely many functions with the same derivative $F' = f$. Their graphs are easy to plot because they differ by a vertical shift:



So, even if we can recover the function F from its derivative F', there are many others with the same derivative, such as $G = F + C$ for any constant real number C. Are there others? Not according to the theorem.

Warning!

It's only true when the domain is an interval.

2.2. Integration over addition and constant multiple: linearity

In this chapter, we will be taking a broader look at how we compute the total value of a function.

For the procedure described in the definition (left), we now have several shortcuts (right):

$$\begin{tikzcd}
\text{sequence} \arrow[d, "RS"] & \text{function on interval} \arrow[l, "\text{sampling}"'] \arrow[d, "\int"] \\
\text{sequence} \arrow[r, "\Delta x\to 0"] & \text{function on interval}
\end{tikzcd}
\qquad
\begin{tikzcd}
x^n \arrow[d, "\int"] & e^x \arrow[d, "\int"] & \sin x \arrow[d, "\int"] & \cos x \arrow[d, "\int"] \\
\frac{1}{n+1}x^{n+1} & e^x & -\cos x & \sin x
\end{tikzcd}$$

Let's review the details.

Suppose we have a partition of interval $[a,b]$. These are the primary nodes:

$$a = x_0,\ \dots,\ x_n = b\,.$$

These are the secondary nodes:

$$c_1,\ \dots,\ c_n\,.$$

They satisfy:

$$x_k \le c_k \le x_{k+1}\,.$$

We start with a function g defined on the secondary nodes. It is simply a sequence of numbers. Here is its *sum*:

$$\Sigma g(x_k) = \sum_{k=1}^{m} g(c_k)\,.$$

It is also defined on the partition. On the one hand, it is simply a sequence of numbers. On the other, it's a function defined on the primary nodes.

What this means is that this procedure is a special kind of function, a function of functions:

$$\text{function} \quad \to \quad \boxed{\quad \Sigma\square \quad} \quad \to \quad \text{another function, sequence}$$

Now we have another function f defined on the secondary nodes. To find the *Riemann sum* Σ of f, we just apply the sum construction to f with its values multiplied by $\Delta x = x_{k+1} - x_k$. We have created another function of functions:

$$\text{function} \quad \to \quad \boxed{\quad \Sigma\square \cdot \Delta x \quad} \quad \to \quad \text{another function, sequence}$$

Next, the *Riemann integral* is defined as a limit. It is the limit of the Riemann sums of ff over all partitions of the interval. This process creates a special kind of function too, also a function of functions:

$$\text{function} \quad \to \quad \boxed{\quad \int \square\, dx \quad} \quad \to \quad \text{another function}$$

As a reminder, the motion analogy for these three operations is as follows:

$$\begin{array}{c}
\text{velocity} \\
\downarrow{\scriptstyle\text{sampling}} \\
\text{sampled velocity} \\
\downarrow{\scriptstyle\cdot\Delta} \\
\text{sampled displacements} \\
\downarrow{\scriptstyle\Sigma} \\
\text{sampled total displacement} \\
\downarrow{\scriptstyle\Delta x\to 0} \\
\text{total displacement}
\end{array}$$

We need to understand how these three functions operate.

We consider the *algebraic properties* of the sums, the Riemann sums, and the Riemann integrals. In light of the Fundamental Theorem of Calculus, we can predict that they will match the algebraic properties of the differences, the difference quotients, and the derivatives.

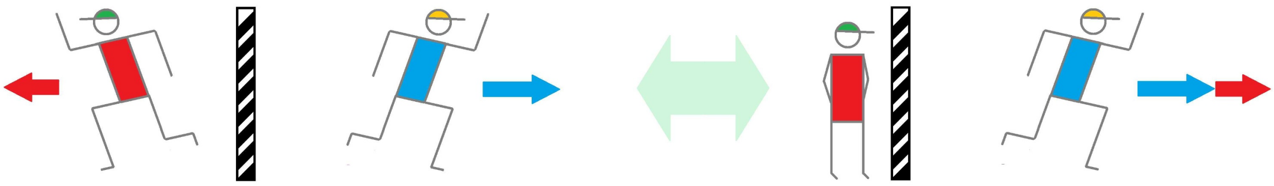
Furthermore, the last item on either list comes from the algebraic properties of the limits. In fact, the ideas of what shortcuts to look for come from those for *limits*: the Sum Rule, the Constant Multiple Rule, etc. The question we will be asking is:

- ▶ What happens to the output function of integration as we perform algebraic operations with the input functions?

There are a few shortcut properties.

Let's take another look at this elementary statement about *motion*:

- ▶ `IF` two runners are running away from a post, `THEN` the distance between them is the same as the distance from the post of a third person running for the both of them with the combined speed.



When two functions are added, what happens to their sums?

This simple algebra, the *Associative Property*, tells the whole story:

$$\begin{array}{rccc} & u & + & U \\ + & & & \\ & v & + & V \\ \hline = & (u+v) & + & (U+V) \end{array}$$

The rule applies even if we have more than just two terms; it's just re-arranging the terms:

$$\begin{array}{rcl} u_p & + & U_p \\ u_{p+1} & + & U_{p+1} \\ \vdots & \vdots & \vdots \\ u_q & + & U_q \\ \hline u_p + ... + u_q & + & U_p + ... + U_q \\ = (u_p + U_p) + & ... & +(u_q + U_q) \\ = \sum_{n=p}^{q} (u_n & + & U_n) \end{array}$$

That's the *Sum Rule for Sums of Sequences* (Chapter 1PC-1). We restate it for functions defined on partitions:

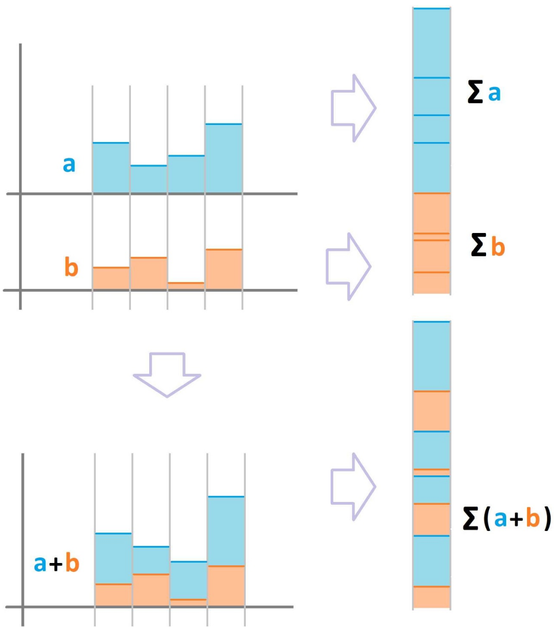
Theorem 2.2.1: Sum Rule for Sums

For any partition of an interval, the sum of the sum of two functions is the sum of the sums of the functions.

In other words, we have:

$$\boxed{\Sigma\,(f+g) = \Sigma f + \Sigma g}$$

For the Riemann sums, the interpretation is also simple; the picture below illustrates how adding two functions causes adding the areas under their graphs:



Theorem 2.2.2: Sum Rule for Riemann Sums

For any partition of an interval, the Riemann sum of the sum of two functions is the sum of the Riemann sums of the functions.

In other words, we have:

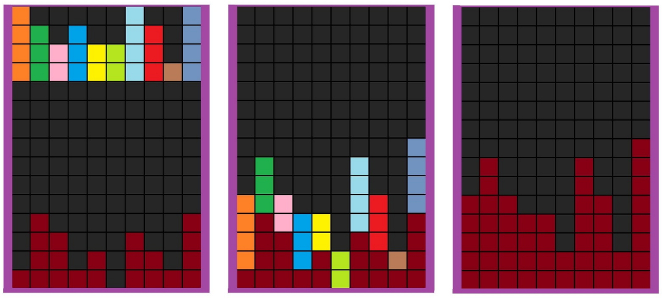
$$\boxed{\Sigma\,(f+g)\cdot\Delta x=\Sigma f\cdot\Delta x+\Sigma g\cdot\Delta x}$$

Proof.

Applying the definition to the function $f+g$, we have:

$$\begin{aligned}\Sigma\big(f+g\big)\cdot\Delta x\ &=\big(f(a)+g(a)\big)+\big(f(a+h)+g(a+h)\big)+...\\&\quad+\big(f(x-h)+g(x-h)\big)\cdot\Delta x\\&=\bigg(f(a)+f(a+h)+f(a+2h)+...+f(x-h)+f(x)\bigg)\cdot\Delta x\\&\quad+\bigg(g(a)+g(a+h)+g(a+2h)+...+g(x-h)+g(x)\bigg)\cdot\Delta x\\&=\Sigma f\cdot\Delta x\,(b)+\Sigma g\cdot\Delta x\,.\end{aligned}$$

It is as if one makes two consecutive drops in a game of Tetris:



Now $\Delta x\to 0$ in these Riemann sums:

Theorem 2.2.3: Sum Rule for Integrals

The integral of the sum of two functions is the sum of the integrals of the functions.

In other words, if f and g are integrable functions over $[a,b]$, then so is $f+g$ and we have:

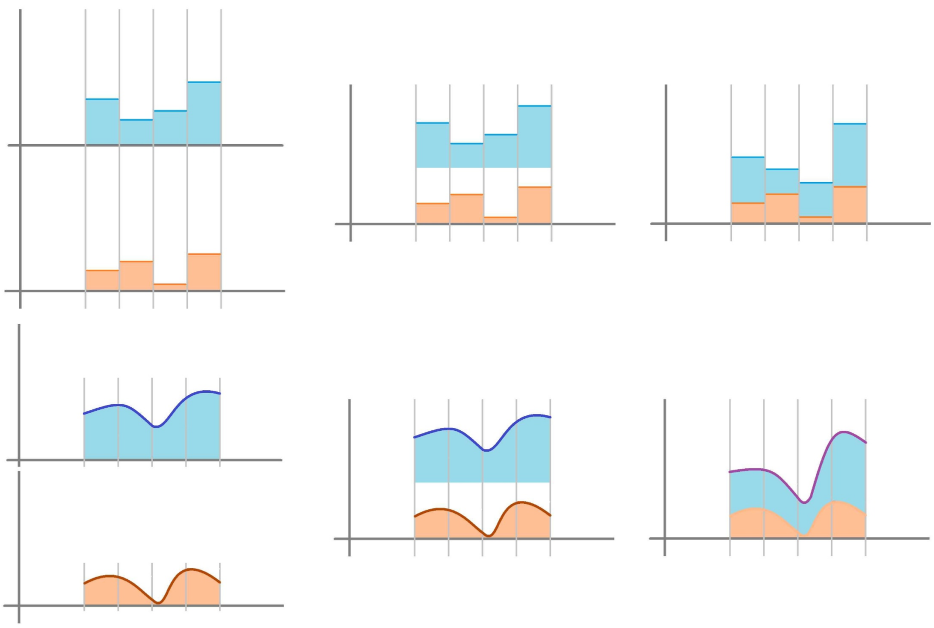
$$\boxed{\int_a^x (f+g)\,dx = \int_a^x f\,dx + \int_a^x g\,dx}$$

Proof.

We use the *Sum Rule for Limits* from Chapter 2DC-1.

So, Δ's become d's!

The picture below illustrates what happens when the bottom drops from a bucket of sand and it falls on a uneven surface:



The last two theorems demonstrate that this is true whether the surface is staircase-like or curved.

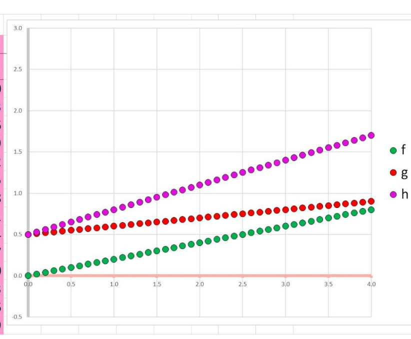
Exercise 2.2.4

Modify the proof for: (a) left-end, (b) mid-point, and (c) general Riemann sums.

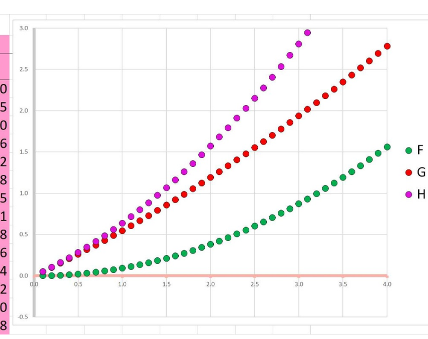
Computationally, we have the same idea as the one for the derivatives: *The integral is split in half.*

Either of the last two theorems can also be demonstrated with the spreadsheet:

n=	40				
h=	0.10				
		x	f	g	h=f+g
	1	0.0	0.00	0.50	0.50
	2	0.1	0.02	0.51	0.53
	3	0.2	0.04	0.52	0.56
	4	0.3	0.06	0.53	0.59
	5	0.4	0.08	0.54	0.62
	6	0.5	0.10	0.55	0.65
	7	0.6	0.12	0.56	0.68
	8	0.7	0.14	0.57	0.71
	9	0.8	0.16	0.58	0.74
	10	0.9	0.18	0.59	0.77
	11	1.0	0.20	0.60	0.80
	12	1.1	0.22	0.61	0.83
	13	1.2	0.24	0.62	0.86
	14	1.3	0.26	0.63	0.89



n=	40				
h=	0.10				
		x	∫f	∫g	∫h
	1	0.0	0.00	0.00	0.00
	2	0.1	0.00	0.05	0.05
	3	0.2	0.00	0.10	0.10
	4	0.3	0.01	0.15	0.16
	5	0.4	0.01	0.21	0.22
	6	0.5	0.02	0.26	0.28
	7	0.6	0.03	0.32	0.35
	8	0.7	0.04	0.37	0.41
	9	0.8	0.06	0.43	0.48
	10	0.9	0.07	0.49	0.56
	11	1.0	0.09	0.55	0.64
	12	1.1	0.11	0.61	0.72
	13	1.2	0.13	0.67	0.80
	14	1.3	0.16	0.73	0.88



The same proof applies to *subtraction* of the sums.

Exercise 2.2.5

State the Difference Rule.

Next, when a function is multiplied by a constant, what happens to its sums?

This simple algebra, the *Distributive Property*, tells the whole story:

$$\begin{array}{lccc} c\cdot(& u & + & U) \\ = & cu & + & cU \end{array}$$

The rule applies even if we have more than just two terms; it's just *factoring*:

$$\begin{array}{rcl} c & \cdot & u_p \\ c & \cdot & u_{p+1} \\ \vdots & \vdots & \vdots \\ c & \cdot & u_q \\ \hline c & \cdot & (u_p + ... + u_q) \\ = c & \cdot & \sum_{n=p}^{q} u_n \end{array}$$

That's the *Constant Multiple for Sums* for sequences (Chapter 1PC-1). We restate it for functions defined on partitions:

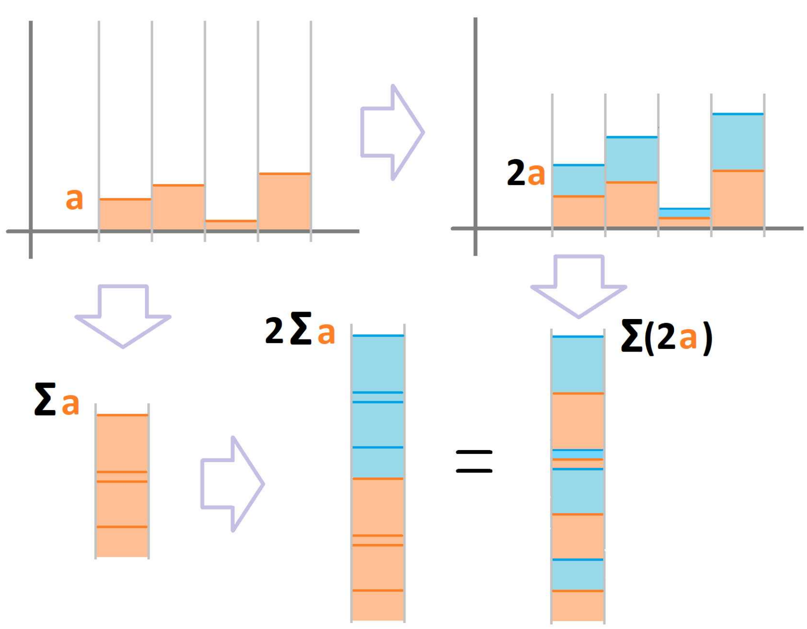
Theorem 2.2.6: Constant Multiple Rule for Sums

For any partition of an interval, the sum of a multiple of a function is the multiple of its sum.

In other words, we have:

$$\boxed{\Sigma(cf) = c\,(\Sigma f)}$$

For the Riemann sums, the interpretation is also simple; the picture below illustrates the idea of multiplication of the function, viz. multiplication of the area under its graph:



Theorem 2.2.7: Constant Multiple Rule for Riemann Sums

The Riemann sum of a multiple of a function is the multiple of its Riemann sum.

In other words, for any real c we have:

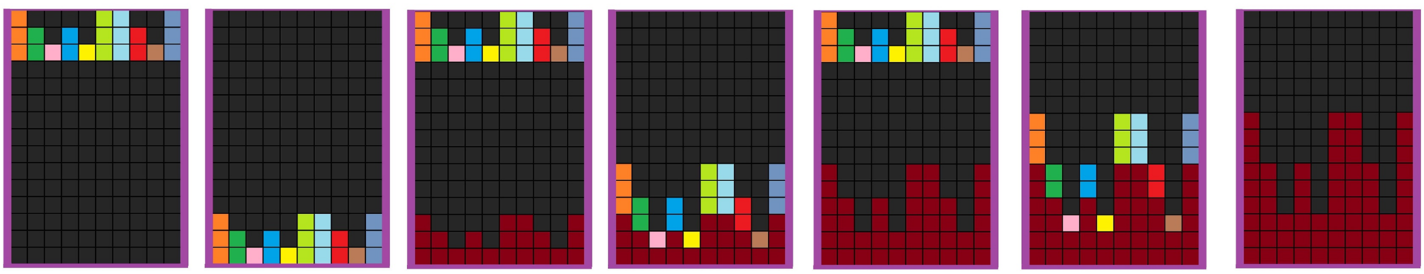
$$\boxed{\Sigma(cf)\cdot\Delta x = c\,(\Sigma f\cdot\Delta x)}$$

Proof.

Applying the definition to the function f, we have:

$$\begin{aligned}\Sigma(cf)\Delta x &= cf(a)+cf(a+h)+cf(a+2h)+...+cf(x-h)\\ &= c\,(f(a)+f(a+h)+f(a+2h)+...+f(x-h))\\ &= c\,\Sigma f\cdot\Delta x\,.\end{aligned}$$

It is as if one is makes several identical drops in a game of Tetris:



Now $\Delta x\to 0$ in these Riemann sums:

Theorem 2.2.8: Constant Multiple Rule for Integrals

The integral of a multiple of a function is the multiple of its integral.

In other words, if f is an integrable function over $[a,b]$, then so is $c\cdot f$ for any real c, and we have:

$$\boxed{\int_a^x (cf)\,dx = c\int_a^x f\,dx}$$

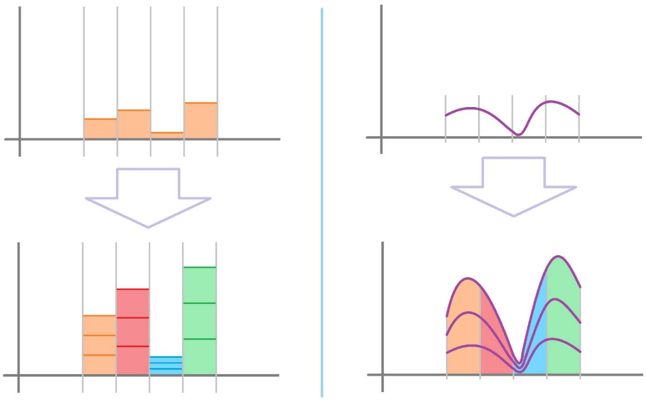
Proof.

We use the *Constant Multiple Rule for Limits* from Chapter 2DC-1.

And Δ's become d's again.

Computationally, we have the same idea as the one for the derivatives: *The constant multiple is factored out of the integral.*

The picture below illustrates the idea that tripling the height of a road will need tripling the amount of soil under it:

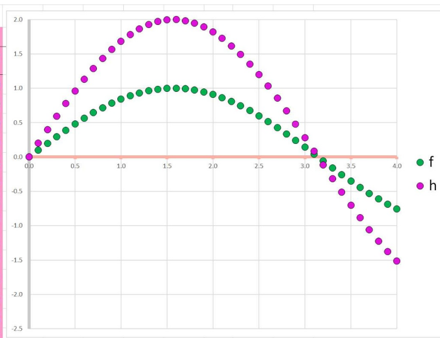


The two last theorems demonstrate that this is true whether the surface is staircase-like or curved.

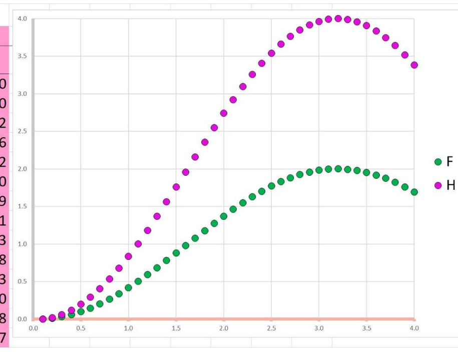
For the *motion* metaphor, if your velocity is tripled, then so is the distance you have covered over the same period of time.

Either of the two last theorems can be demonstrated with the spreadsheet:

n=	40			
h=	0.10		2.00	
		x	f	h=2f
	1	0.0	0.00	0.00
	2	0.1	0.10	0.20
	3	0.2	0.20	0.40
	4	0.3	0.30	0.59
	5	0.4	0.39	0.78
	6	0.5	0.48	0.96
	7	0.6	0.56	1.13
	8	0.7	0.64	1.29
	9	0.8	0.72	1.43
	10	0.9	0.78	1.57
	11	1.0	0.84	1.68
	12	1.1	0.89	1.78
	13	1.2	0.93	1.86
	14	1.3	0.96	1.93



n=	40			
h=	0.10			
		x	∫f	∫h
	1	0.0	0.00	0.00
	2	0.1	0.00	0.00
	3	0.2	0.01	0.02
	4	0.3	0.03	0.06
	5	0.4	0.06	0.12
	6	0.5	0.10	0.20
	7	0.6	0.15	0.29
	8	0.7	0.20	0.41
	9	0.8	0.27	0.53
	10	0.9	0.34	0.68
	11	1.0	0.42	0.83
	12	1.1	0.50	1.00
	13	1.2	0.59	1.18
	14	1.3	0.68	1.37



Exercise 2.2.9

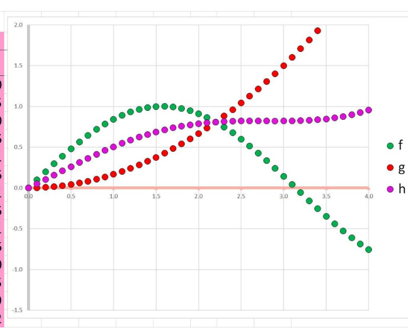
Modify the proof for: (a) left-end, (b) mid-point, and (c) general Riemann sums.

These two operations can be combined into one producing *linear combinations*:

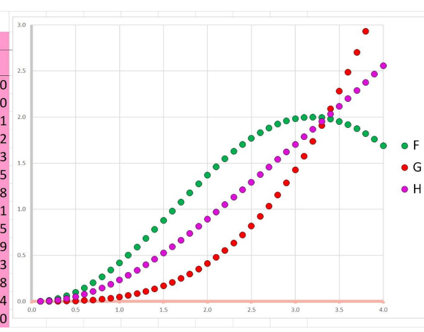
$$\alpha x + \beta y\,,$$

where α, β are two constant numbers. The idea applies to functions; for example, this is the average of two functions (left):

n=	40				
h=	0.10		0.50	0.50	
		x	f	g	h
	1	0.0	0.00	0.00	0.00
	2	0.1	0.10	0.00	0.05
	3	0.2	0.20	0.01	0.10
	4	0.3	0.30	0.02	0.16
	5	0.4	0.39	0.03	0.21
	6	0.5	0.48	0.04	0.26
	7	0.6	0.56	0.06	0.31
	8	0.7	0.64	0.08	0.36
	9	0.8	0.72	0.11	0.41
	10	0.9	0.78	0.14	0.46
	11	1.0	0.84	0.17	0.50
	12	1.1	0.89	0.20	0.55
	13	1.2	0.93	0.24	0.59
	14	1.3	0.96	0.28	0.62



n=	40				
h=	0.10				
		x	∫f	∫g	∫h
	1	0.0	0.00	0.00	0.00
	2	0.1	0.00	0.00	0.00
	3	0.2	0.01	0.00	0.01
	4	0.3	0.03	0.00	0.02
	5	0.4	0.06	0.00	0.03
	6	0.5	0.10	0.01	0.05
	7	0.6	0.15	0.01	0.08
	8	0.7	0.20	0.02	0.11
	9	0.8	0.27	0.02	0.15
	10	0.9	0.34	0.03	0.19
	11	1.0	0.42	0.05	0.23
	12	1.1	0.50	0.06	0.28
	13	1.2	0.59	0.08	0.34
	14	1.3	0.68	0.11	0.40



We also notice what happens to their integrals (right):

- ▶ The integral of the average is the average of the integrals.

The question becomes: What happens to linear combinations of functions under integration?

Recall that a function F is linear if it "preserves" linear combinations:

$$\alpha x + \beta y \ \to\ \boxed{\quad F \quad} \ \to\ \alpha F(x) + \beta F(y)$$

With this idea, these two formulas can be combined into one. The sum, the Riemann sum, and the Riemann integral are linear functions of functions. A precise version is below:

Theorem 2.2.10: Linearity of Integration

The sum, the Riemann sum, and the Riemann integral of a linear combination of two functions is the linear combination of their sums, Riemann sums, and Riemann integrals, respectively, whenever they exist.

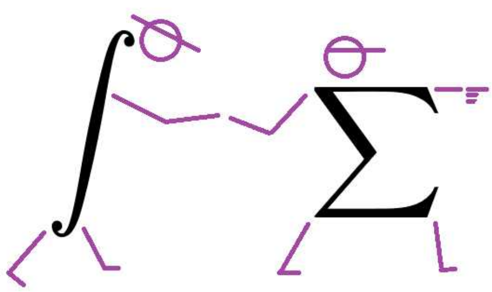
In other words, we have:

$$\boxed{\begin{array}{lll}\Sigma(\alpha f+\beta g) & =\alpha\Sigma f & +\beta\Sigma g\\[2ex] \Sigma(\alpha f+\beta g)\cdot\Delta x & =\alpha\Sigma f\cdot\Delta x & +\beta\Sigma g\cdot\Delta x\\[1ex] \displaystyle\int_a^x(\alpha f+\beta g)\,dx & =\alpha\displaystyle\int_a^x f\,dx & +\beta\displaystyle\int_a^x g\,dx\end{array}}$$

The last formula is illustrated with the following diagram:

$$\alpha f+\beta g\ \to\ \boxed{\int}\ \to\ \alpha\int f\,dx+\beta\int g\,dx$$

As we see, the integral follows the Riemann sum, every time:



In summary, we take the shortcut in our diagram on many occasions and ignore the rest:

$$\begin{array}{ccccc}\text{sequence} & \xleftarrow{\text{sampling}} & \text{function on }[a,b] & & \\ \downarrow{\scriptstyle\text{RS}} & & \downarrow{\scriptstyle\text{integral}} & \longrightarrow & \boxed{\begin{array}{c}f\\ \downarrow{\scriptstyle\int}\\ \displaystyle\int f\,dx\end{array}}\\ \text{sequence} & \xrightarrow{\Delta x\to 0} & \text{function on }[a,b] & & \end{array}$$

Warning!

These rules can be acquired via the Fundamental Theorem of Calculus.

2.3. *Integration in light of the Fundamental Theorem*

Let's review what we know about anti-differentiation.

Definition 2.3.1: antiderivative on partition

Suppose a function f is defined on the secondary nodes of a partition of an interval I. Then a function F defined on the nodes of the partition that satisfies the equation:

$$\boxed{\frac{\Delta F}{\Delta x}(c) = f(c)}$$

for every secondary node c, is called an *antiderivative* of f.

In Volume 1, we found a recursive solution of this equation by solving it for the difference quotient as in the case of acquiring the position from the velocity:

$$v_n = \frac{p_{n+1} - p_n}{\Delta t} \implies p_{n+1} = p_n + v_n \Delta t \,.$$

It's pure algebra!

The continuous case is by far more complex:

Definition 2.3.2: antiderivative on interval

Suppose a function f is defined on an interval I. Then a differentiable function F defined on I that satisfies the equation:

$$\boxed{\frac{dF}{dx}(x) = f(x)}$$

for every x, is called an *antiderivative* of f.

We use "an" because there may be many antiderivatives for each function.

We can think of the definition as an *equation*, an equation for functions:

Given f, solve for F	Given f, solve for F
$\frac{\Delta F}{\Delta x} = f$	$\frac{dF}{dx} = f$

This equation has infinitely many solutions when f is integrable. Furthermore, according to the *Antidifferentiation Theorem*, if F is one of its antiderivatives, then the set of all antiderivatives of f is

$$\{F + C : \ C \text{ real }\}\,.$$

The Riemann integrals are linked to antiderivatives by the following:

Theorem 2.3.3: Fundamental Theorem of Calculus I and II

I *For any continuous function f on $[a,b]$, the function defined by*

$$F(x) = \int_a^x f\,dx$$

is an antiderivative of f.

II *For any integrable function f on $[a,b]$ and any of its antiderivatives F, we*

have

$$\int_a^b f\, dx = F(b) - F(a)\,.$$

Why *two* parts? Because we deal with *two* operations of calculus – differentiation and integration – and we can compose them in *two* ways. According to the Fundamental Theorem, the operations of differentiation and integration cancel each other as follows:

$$\begin{array}{lccccccccc}
\text{FTCI:} & f & \to & \boxed{\int \square\, dx} & \to & F & \to & \boxed{\frac{d}{dx}\square} & \to & f \\
\text{FTCII:} & F & \to & \boxed{\frac{d}{dx}\square} & \to & f & \to & \boxed{\int \square\, dx} & \to & F+C
\end{array}$$

The properties of the Riemann integrals are proven *from scratch* in the last section. In this section, we acquire the same results by applying the Fundamental Theorem to the corresponding theorems about the derivatives (Volume 2).

We start over.

We can restate the *Anti-differentiation Theorem* as follows:

Corollary 2.3.4: Set of Antiderivatives

Suppose F is any antiderivative of a function f defined on an interval. Then the set of all of its antiderivatives is

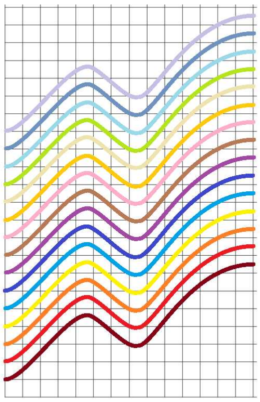
$$\{F + C :\ C \text{ real}\}\,.$$

As it is often the case with equations, there seems to be many (infinitely many) solutions. But a very important conclusion is that *it suffices to find just one anti-derivative!*

Warning!

The formula $F + C$ works only when the function is defined on an interval.

This is what this set of functions would look like:



This is the solution set of the equation and it may be called *the* antiderivative.

Exercise 2.3.5: x^3

Find all F that satisfy the equation:

$$F'(x) = x^3\,.$$

Exercise 2.3.6

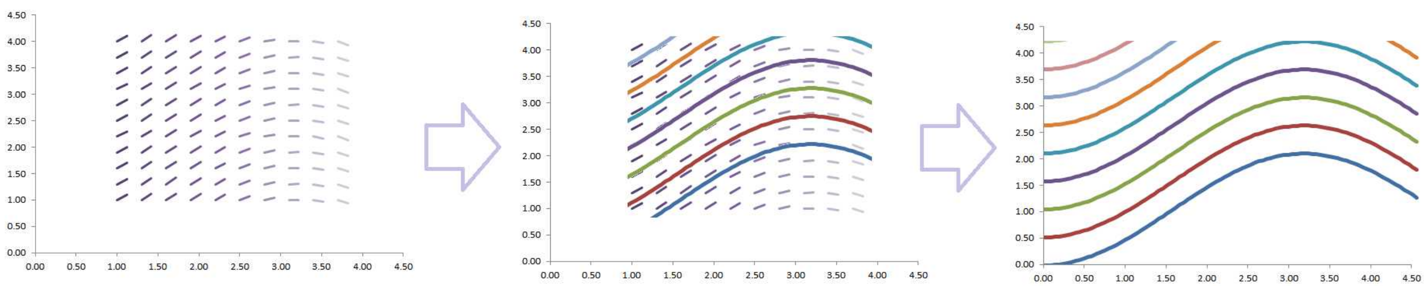
Suppose a function f is defined on an open interval I. Prove that the following:

1. The graphs of two different antiderivatives of f never intersect.
2. For every point (x,y) with x within I, there is an antiderivative of f the graph of which passes through it.

The problem then becomes the one of finding a single function F, either

- from its difference quotient $\frac{\Delta F}{\Delta x}$, or
- from its derivative $\frac{dF}{dx}$.

In other words, we reconstruct the function from a "field of slopes":

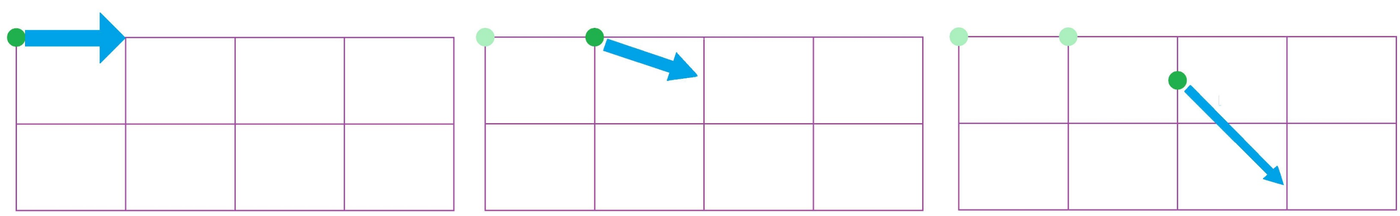


One can imagine a flowing liquid with its direction known at every location. How do we find the path of a particular particle? The process of reconstructing a function, F, from its derivative, f, is called anti-differentiation or *integration*.

The integration problem has been solved on several occasions for the former, discrete case – velocity from acceleration and location from velocity – via these recursive formulas:

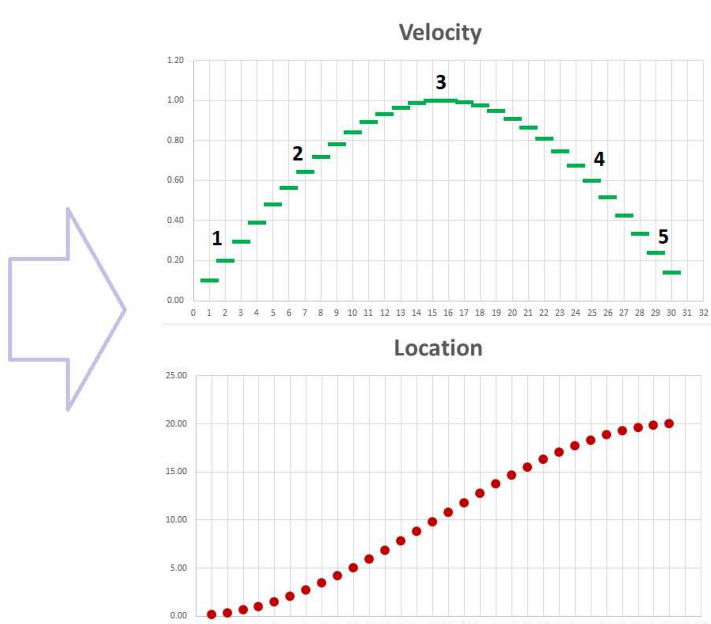
$$F(x_{n+1}) = F(x_n) + f(c_n)\Delta x_n \,.$$

For each location, we look up the velocity, find the next location, and repeat:



If the nodes of the partition are close enough to each other, these points form curves:

time	velocity	location
min	miles/min	miles
0		0.00
1	0.10	0.10
2	0.20	0.30
3	0.30	0.59
4	0.39	0.98
5	0.48	1.46
6	0.56	2.03
7	0.64	2.67
8	0.72	3.39
9	0.78	4.17
10	0.84	5.01
11	0.89	5.91
12	0.93	6.84
13	0.96	7.80
14	0.99	8.79
15	1.00	9.78
16	1.00	10.78
17	0.99	11.77
18	0.97	12.75
19	0.95	13.70
20	0.91	14.60
21	0.86	15.47
22	0.81	16.28
23	0.75	17.02
24	0.68	17.70
25	0.60	18.30
26	0.52	18.81
27	0.43	19.24
28	0.33	19.57
29	0.24	19.81
30	0.14	19.95



For the latter, continuous case, this is a challenging problem: How does one plot a curve that follows these – infinitely many – tangents?

To begin with, we just try to *reverse differentiation*. We will try to construct a theory of anti-differentiation that matches – to the degree possible – that of differentiation.

Here is a short *list of derivatives* of functions (for all x for which the function is differentiable):

function	$\longrightarrow$	derivative
x^r		rx^{r-1}
$\ln x$		$\dfrac{1}{x}$
e^x		e^x
$\sin x$		$\cos x$
$\cos x$		$-\sin x$
antiderivative	$\longleftarrow$	function

To find antiderivatives, i.e., integrals, reverse the order:

▶ Read each line from right to left!

Example 2.3.7: $\sin$ and $\cos$

What is an antiderivative of $\cos x$? We need to solve for F:

$$F'(x) = \cos x\,.$$

Just find $\cos x$ in the right column of the table. The corresponding function on the left is $\sin x$. That's the answer: $F(x) = \sin x$!

What is an antiderivative of $\sin x$? Solve:

$$F'(x) = \sin x\,.$$

Just find $\sin x$ in the right column. It's not there... but $-\sin x$ is! The corresponding function on the left is $\cos x$. Then (according to the Constant Multiple Rule) the solution must be $-\sin x$.

It's that simple! We may need some tweaking to make the formulas about to emerge as easy to apply as the original ones.

Example 2.3.8: power formula

For example, let's find antiderivatives of x^n. Use the *Power Formula* for differentiation (the first row), divide by r, and apply the *Constant Multiple Rule*:

$$(x^r)' = rx^{r-1} \implies \frac{1}{r}(x^r)' = x^{r-1} \implies \left(\frac{1}{r}x^r\right)' = x^{r-1}\,.$$

We then simplify the right-hand side by setting $r - 1 = s$:

$$\left(\frac{1}{s+1}x^{s+1}\right)' = x^s\,.$$

We make the right-hand side the left-hand side and we have the *Power Formula* for anti-differentiation. We have the following:

$$\text{An antiderivative of } x^s \text{ is } \tfrac{1}{s+1}x^{s+1}, \text{ provided } s \neq -1.$$

But what if $s = -1$? Then we read the answer from the next line in the table:

$$\text{An antiderivative of } x^{-1} \text{ is } \ln|x|.$$

So, the rule that "the derivative of a power function is a power function of degree 1 lower" has an exception, the 0-power, and the rule that "the antiderivative of a power function is a power function of degree 1 higher" has an exception too.

Taking the rest of these rows, we have a *list of integrals* of functions, on open intervals:

function	$\longrightarrow$	antiderivative/integral
x^s		$\frac{1}{s+1}x^{s+1}, \quad s \neq -1$
$\dfrac{1}{x}$		$\ln\|x\|$
e^x		e^x
$\sin x$		$-\cos x$
$\cos x$		$\sin x$
derivative	$\longleftarrow$	function

Example 2.3.9: domains

Each formula is only valid on an open interval on which the antiderivative is defined. For example, we interpret the second row as follows:

- $\ln(x)$ is an antiderivative of $\frac{1}{x}$ on the interval $(0, +\infty)$, and
- $\ln(-x)$ is an antiderivative of $\frac{1}{x}$ on the interval $(-\infty, 0)$.

Next, we will need the *rules of integration.*

First, consider the *Sum Rule for Derivatives* (Volume 2): The derivative of the sum is the sum of the derivatives; i.e.,

$$(f+g)' = f' + g'\,.$$

Let's read that formula from right to left:

Theorem 2.3.10: Sum Rule for Integrals

An integral of the sum is the sum of the integrals.

In other words, we have:

If

F *is the integral of* f *and*

G *is the integral of* g,

then

$F+G$ *is the integral of* $f+g$.

Proof.

We apply the *Sum Rule for Derivatives* to confirm:

$$(F(x)+G(x))'=F'(x)+G'(x)=f(x)+g(x)\,.$$

Exercise 2.3.11

What about the converse?

Example 2.3.12: sums

Solve for F:

$$F'(x)=x^2+\sin x\,.$$

The equation is solved by solving the following two equations:

$$\text{for } \ G:\ G'(x)=x^2 \ \text{ and for } \ H:\ H'(x)=\sin x\,.$$

The solutions are found in the table:

$$G(x)=\frac{1}{3}x^3 \ \text{ and } \ H(x)=-\cos x\,.$$

According to the theorem, we have the answer on $(-\infty,+\infty)$:

$$F(x)=\frac{1}{3}x^3-\cos x+C\,.$$

Exercise 2.3.13

Using this rule, find the integral of $\ln x^2$.

Compare:

The	derivative	of the sum is the sum of the	derivatives.
The	integral	of the sum is the sum of the	integrals.

Similarly, consider the *Constant Multiple Rule for Derivatives* (Volume 2): The derivative of a multiple is the multiple of the derivative; i.e.,

$$(cf)'=cf'\,.$$

Let's read that formula from right to left:

Theorem 2.3.14: Constant Multiple Rule for Integrals

The integral of a multiple is the multiple of the integral.

In other words, we have:

If

F is the integral of f and

c is a constant,

then

cF is the integral of cf.

Proof.

We apply the *Constant Multiple Rule for Derivatives*:

$$(cF(x))' = cF'(x) = cf(x)\,.$$

Exercise 2.3.15

What about the converse?

Example 2.3.16: constant multiples

Solve for F:

$$F'(x) = 3\sin x\,.$$

We solve the equation by solving the following equation:

$$\text{for } \ G:\ G'(x) = \sin x\,.$$

The solution is found in the table:

$$G(x) = -\cos x\,.$$

According to the theorem, we have an answer on $(-\infty,+\infty)$:

$$F(x) = 3(-\cos x) + C\,.$$

Exercise 2.3.17

Using this rule, find the integral of e^{x+3}.

Compare:

The	derivative	of a constant multiple is the constant multiple of the	derivative.
The	integral	of a constant multiple is the constant multiple of the	integral.

As we know from the *Anti-differentiation Theorem*, every antiderivative of a function comes with infinitely many others:

$$F \ \to \ F + C \ \text{ for every real } C\,,$$

on every open interval. Together they form *the* antiderivative or the integral of the function.

Example 2.3.18: free fall

We can make our analysis of *free fall* more specific:

Functions of time		
The acceleration is constant.	$a = -g$	$\Longrightarrow$
The velocity is linear.	$v = -gt + C$	$\Longrightarrow$
The location is quadratic.	$p = -gt^2/2 + Ct + K$	

The constants C and K cover all possible trips of the ball.

This is how we rewrite the above list:

$$\begin{aligned}
\int x^s\,dx &= \frac{1}{s+1}x^{s+1}+C\,, \quad \text{for } s\neq -1\\
\int \frac{1}{x}\,dx &= \ln x + C\\
\int e^x\,dx &= e^x + C\\
\int \sin x\,dx &= -\cos x + C\\
\int \cos x\,dx &= \sin x + C
\end{aligned}$$

We restate the rules too.

Sum Rule:

$$\int (f+g)\,dx = \int f\,dx + \int g\,dx$$

Constant Multiple Rule:

$$\int (cf)\,dx = c\int f\,dx$$

Exercise 2.3.19

State the Linear Composition Rule for integrals.

With these rules, when applicable, integration is very similar to differentiation.

Example 2.3.20: rules of integration

Find the integral of

$$3x^2+5e^x+\cos x\,.$$

One can imagine what he'd do to differentiate and then follow the same steps but with the integration formulas and rules used instead.

Differentiation:

$$\begin{aligned}
(3x^2+5e^x+\cos x)' &= (3x^2)'+5(e^x)'+(3\sin x)' && \text{SR}\\
&= 3(x^2)'+5(e^x)'+3(\sin x)' && \text{CMR}\\
&= 3\cdot 6x+5e^x+3\cos x\,. && \text{Table}
\end{aligned}$$

Integration:

$$\begin{aligned}
\int (3x^2+5e^x+\cos x)'dx &= \int (3x^2)\,dx + \int 5(e^x)\,dx + \int (3\sin x)dx && \text{SR}\\
&= 3\int (x^2)\,dx + 5\int (e^x)\,dx + 3\int (\sin x)dx && \text{CMR}\\
&= 3\cdot x^3/3+5e^x+3(-\cos x)+C\,. && \text{Table}
\end{aligned}$$

Just as when solving equations, we can easily confirm that our computations are correct, by substitu-

tion. In this case, we *differentiate the integral*:

$$\begin{aligned}(x^3+5e^x+3\sin x)' &= (x^3)'+5(e^x)'+(3\sin x)'\\ &= 3x^2+5e^x+3\cos x\,.\end{aligned}$$

This is the original function! The answer checks out.

Exercise 2.3.21

Find the integral of $5e^{3x+2}-e^e$.

Below, we have these two diagrams to illustrate the interaction of integrals with algebra:

$$\begin{tikzcd} f,g \arrow[d, "+"] & f',g' \arrow[l, "\int"'] \arrow[d, "+"] \\ f+g & f'+g' \arrow[l, "\int"'] \end{tikzcd} \qquad \begin{tikzcd} f \arrow[d, "\cdot c"] & f' \arrow[l, "\int"'] \arrow[d, "\cdot c"] \\ cf & cf' \arrow[l, "\int"'] \end{tikzcd}$$

The arrows of differentiation are reversed! We start with a pair of functions at top right, then we proceed in two ways:

- Left: integrate them. Then down: add the results.
- Down: add them. Then left: integrate the results.

The result is the same!

So far, this is very similar to differentiation. The strategy is the same: divide and concur. Split addition with the Sum Rule, then factor out the coefficients with the Constant Multiple Rule, then apply the table results to these pieces.

Unfortunately, this is where the similarities stop.

There is no analog of the Product Rule for the Derivatives:

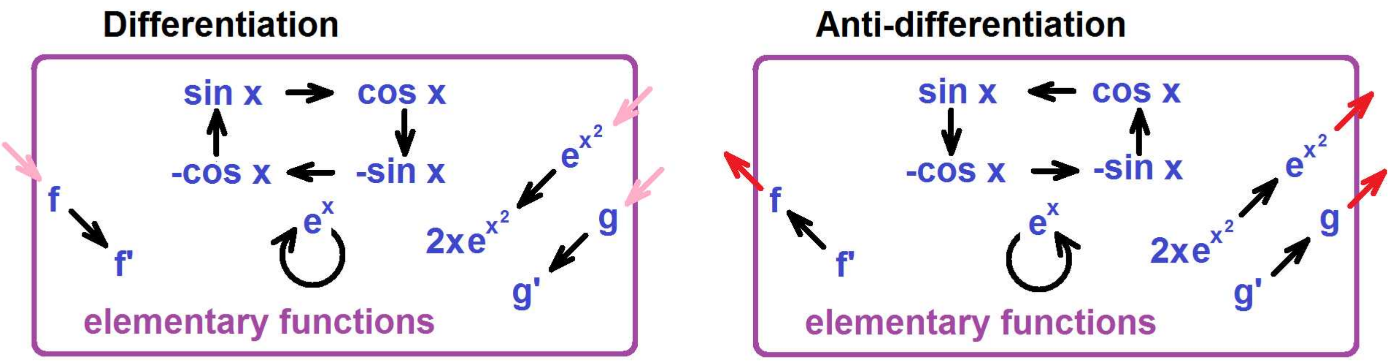
The	derivative	of the product of two functions	*can*	
		be expressed in terms of their	derivatives	and the functions themselves.
The	integral	of the product of two functions	*cannot*	
		be expressed in terms of their	integrals	and the functions themselves.

Similarly there is no Quotient Rule, nor the Chain Rule, for integration.

This difference has profound consequences. We can start with just these functions:

$$x^s,\ \sin x,\ e^x\,.$$

Then – by applying the four algebraic operations, composition, and inverting – we can construct a great variety of functions. Let's call them "elementary functions". Because of the way they are constructed, *all* of them can be easily differentiated with the rules of differentiation, thus producing other elementary functions (left):



However, contrary to what the above list might suggest, integration will often take us outside of the realm of elementary functions (right). For example, a new function, called the *Gauss error function*, must be created for this important integral:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int e^{-x^2}\, dx\,.$$

The result will be the same if we *exclude* from the "elementary functions" either the trigonometric functions or the exponent. The result will be the same if we *include* more functions to the list.

With the help of the *Anti-differentiation Theorem*, we can claim that we have found *all* antiderivatives of these functions on our list, over open intervals within the domains of integral, i.e., *the* integral of the corresponding function:

Definition 2.3.22: general antiderivative and indefinite integral

For a given function f, *the general antiderivative or the indefinite integral of f over open interval I* is defined by:

$$\int f\, dx = F(x) + C\,,$$

where F is any antiderivative of f on I, i.e., $F' = f$, understood as the collection of all such functions over all possible real numbers C. This collection is also called *the indefinite integral* of f.

Example 2.3.23: how many antiderivatives?

There are infinitely many antiderivatives, but there is more to it. Let's take a more careful look at one line on the list:

$$\int \frac{1}{x}\, dx \overset{???}{=\!=} \ln|x| + C, \quad x \neq 0\,.$$

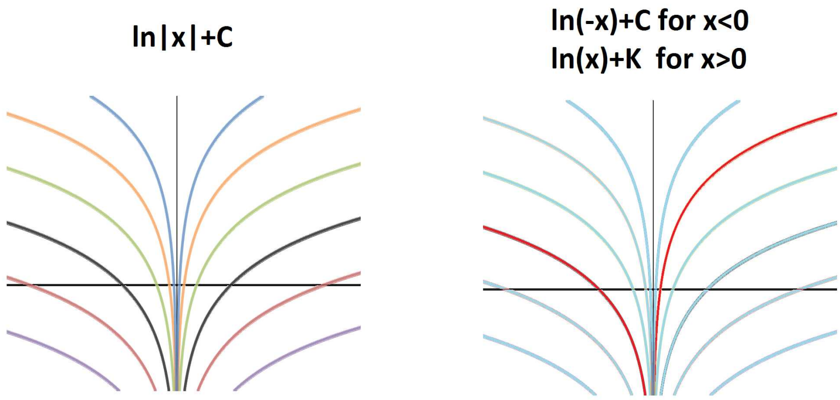
This formula is intended to mean the following:

1. We have captured infinitely many – one for each real number C – antiderivatives.
2. We have captured *all* of them.

However, the *Anti-differentiation Theorem* applies only to *one interval at a time*. Meanwhile, the domain of $1/x$ consists of two rays $(-\infty, 0)$ and $(0, +\infty)$. As a result, we solve this problem *separately* on either of the two intervals. Then the antiderivatives of $1/x$ are:

- $\ln(-x) + C$ on $(-\infty, 0)$, and
- $\ln(x) + C$ on $(0, +\infty)$.

But if now we were to combine each of these pairs of functions into one, F, defined on $(-\infty, 0) \cup (0, +\infty)$, we would realize that, every time, the two constants might be different. After all, they have nothing to do with each other! We illustrate the wrong (incomplete) answer on the left, and the correct one on the right:



The image on the left, as well as the formula we started with, might suggest that all of the function's antiderivatives are even functions. The image on the right shows a single antiderivative (in red) but its two branches don't have to match!

Algebraically, the antiderivative of $\frac{1}{x}$ – on the whole domain – is given by this piecewise-defined function:

$$F(x) = \begin{cases} \ln(-x) + C & \text{for } x \text{ in } (-\infty, 0), \\ \ln(x) + K & \text{for } x \text{ in } (0, +\infty). \end{cases}$$

It has *two* parameters instead of the usual one. The number matches the number of the "components" of the domain.

Exercise 2.3.24

Verify that this is an antiderivative of $1/x$.

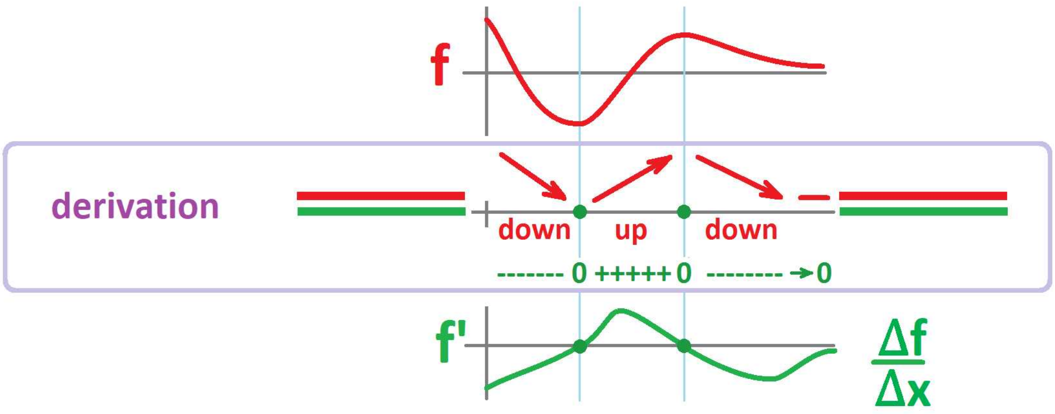
Exercise 2.3.25

In a similar fashion, examine the Power Formula above for $s < -1$.

Example 2.3.26: find graphs

The antiderivatives on our list were discovered by reading the results of differentiation backwards. We can do the same for graphs.

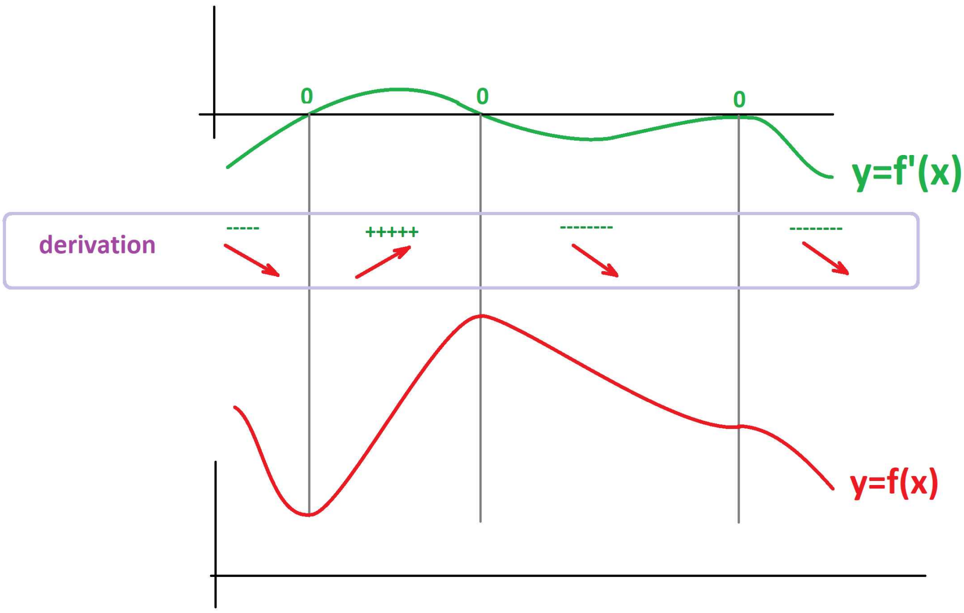
Below, the derivative's graph (green) was found from the graph of the function (red) by looking at the monotonic behavior of f (either $f' > 0$ or $f' < 0$, and local extreme points: $f' = 0$):



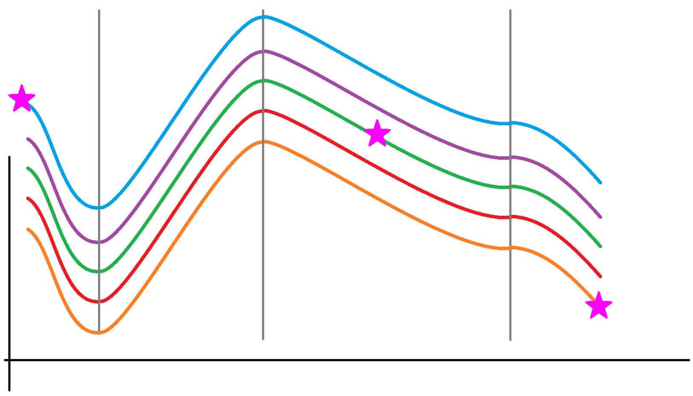
In summary, we look at $\searrow$, $\nearrow$ of f to find $+$, $-$ of f'.

In reverse, we look at $+$, $-$ of f' to find $\searrow$, $\nearrow$ of f.

Here is an example of how the graph of f is found from the graph of its derivative:



To show *the complete answer*, we have to show multiple antiderivatives:



When the initial (or final, or mid-flight) state is known, we pick a specific curve from this set.

Exercise 2.3.27

Find the inflection points.

2.4. Linear change of variables in integral

Numerous quantities are defined via the Riemann sums: areas and displacements as presented in Chapter 1; and weights, volumes, lengths, fluxes, work, and many more to be presented in Chapter 3. The latter part of the Fundamental Theorem allows us to compute anything defined this way by means of a simple substitution – as long as an antiderivative can be found! We will refer to them as *indefinite integrals* or simply integrals.

How do we find integrals? This is the subject of the present chapter.

In the last section, we saw some facts about integrals as they are matched against those of the derivatives. In this section, we will further examine another such fact, the *Linear Composition Rule.*

The variables of the functions we are considering are quantities we meet in everyday life. Frequently, there are multiple ways to measure these quantities:

- length and distance: inches, miles, meters, kilometers, ..., light years

- area: square inches, square miles, ..., acres
- volume: cubic inches, cubic miles, ..., liters, gallons
- time: minutes, seconds, hours, ..., years
- weight: pounds, grams, kilograms, karats
- temperature: degrees of Celsius, of Fahrenheit
- money: dollars, euros, pounds, yen
- etc.

The conversion formulas for these units are seen in mathematics as *changes of variables*.

Almost all conversion formulas are just multiplications, such as this one:

$$\#\ \text{ of meters } = \#\ \text{ of kilometers } \cdot 1000\,.$$

Warning!

We don't convert "pounds to kilos", we convert the *number of* pounds to the *number of* kilos.

Let's consider motion as an example:

- If the distance is measured in *inches* and time in *minutes*, the velocity is measured in *inches per minute*.
- Now, if the distance is measured in *feet*, the velocity is now measured in *feet per minute*.
- But if the time is measured in *seconds*, the velocity is measured in *inches per second*.

We are dealing with the same motion just measured in different units and, therefore, different functions. How do we transition between the three functions?

Suppose a function

$$y = f(x)$$

establishes a relation between two quantities x and y:

$$x \xrightarrow{\quad f \quad} y$$

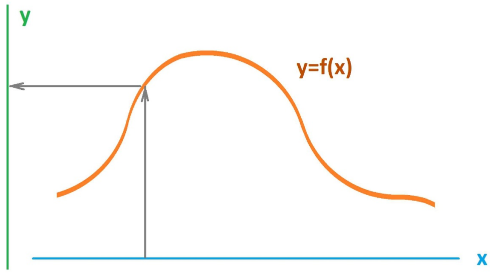
Now, either one may be replaced with a new variable (or a new unit). Let's call them t and z respectively and suppose these replacements, i.e., substitutions, are given by some functions:

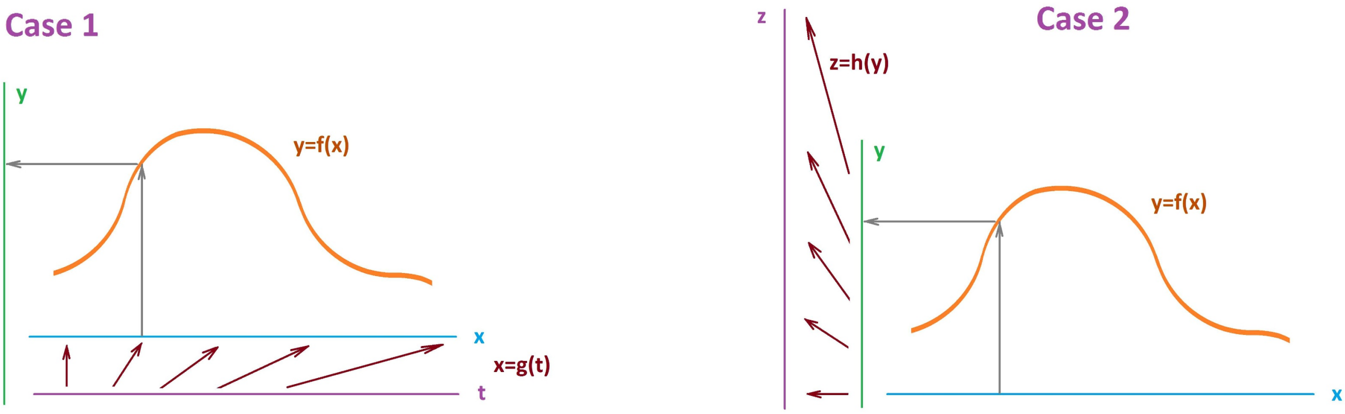
- Case 1: $x = g(t)$ and $y = k(t) = f(g(t))$
- Case 2: $z = h(y)$ and $z = k(x) = h(f(x))$

These substitutions create new relations between the variables:

$$\begin{array}{lccccccc}
\text{Case 1}: & t & \xrightarrow{\quad g \quad} & x & \xrightarrow{\quad f \quad} & y & & \\
\text{Case 2}: & & & x & \xrightarrow{\quad f \quad} & y & \xrightarrow{\quad h \quad} & z
\end{array}$$

The two cases are shown in the picture below:





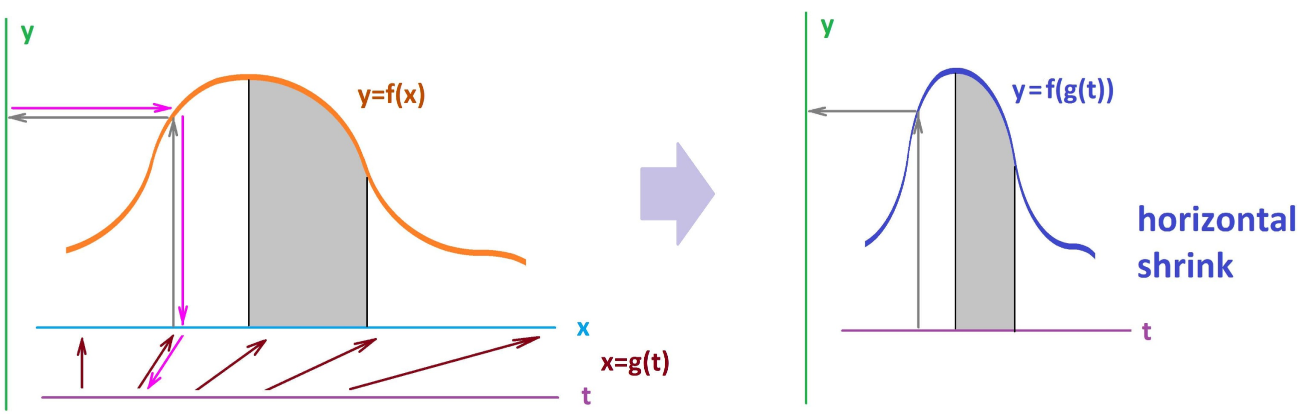
We will assume that these functions g and h are *linear*.

Case 1

The change of units is

$$x = g(t) = mt + b\,.$$

Re-scaling the horizontal axis requires an adjustment of the integral if it is seen as the area under the graph:



Above, the x-axis is shrunk by a factor of 2, i.e., $x = t/2$. Therefore, the integral over any interval in x will have multiplied by 2 to get one for t.

Theorem 2.4.1: Linear Composition Rule I

For any numbers $m \neq 0$ and b and any integrable function f, we have:

$$\boxed{\int f(mt+b)\,dt = \frac{1}{m}\int f(x)\,dx\Big|_{x=mt+b}}$$

Proof.

We take:

$$F(x)=\int f(x)\,dx\,.$$

We apply the *Chain Rule* and the *Constant Multiple Rule for Differentiation*:

$$\frac{d}{dt}\left(\frac{1}{m}F(mt+b)\right)=\frac{1}{m}\frac{d}{dt}\left(F(mt+b)\right)=\frac{1}{m}mF'(mt+b)=F'(mt+b)=f(mt+b)\,.$$

Example 2.4.2: time shift

Suppose x is the time, and suppose we change the moment from which we start measuring time. Then we have:

$$x=t+t_0\ \Longrightarrow\ \int k(x)\,dx=\int f(t+t_0)\,dt\,.$$

Example 2.4.3: seconds to minutes

Suppose x is the time and y is the location. If x is measured in seconds, then switching to time t measured in minutes will require a function:

$$x=g(t)=60t\,.$$

We know from Chapter 2DC-4 that the graphs of the quantities describing motion are simply re-scaled versions of the old ones. Let's recast this statement in the integral form.

- Suppose $y=q(t)$ and $y=p(x)$ are the location as functions of minutes and seconds respectively. Then
$$q(t)=p(60t)\,.$$
- Suppose $v(t)=q'(t)$ and $e(x)=p'(x)$ are the velocities as functions of minutes and seconds respectively. Therefore, $\displaystyle\int v\,dt=q$ and $\displaystyle\int e\,dx=p$. We substitute these into the above equation:

$$\int v\,dt=\int e\,dx\Bigg|_{x=60t}\,.$$

Exercise 2.4.4

Express the location as a function of minutes in terms of the velocity as a function of seconds.

Example 2.4.5: compare to Chain Rule

Let's find both the derivative and the integral of the following function:

$$k(t)=\sin(3t-1)\,.$$

After all, the main challenge might be the decomposition of the function:

$$t\mapsto 3t-1=x\mapsto \sin x=z$$

Here, x is the new variable that we have made up.

The derivative:

$$\frac{d}{dt}k(t)=\frac{d}{dt}\big(\sin(3t-1)\big)=3\frac{d}{dx}\sin x\Big|_{x=3t-1}=3\cos x\Big|_{x=3t-1}=3\cos(3t-1)\,.$$

The integral:

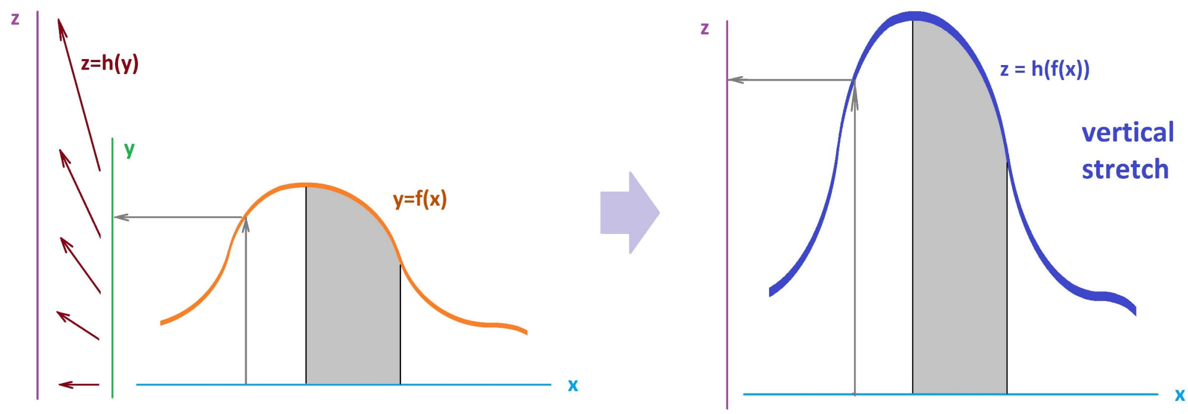
$$\int k(t)\,dt=\frac{1}{3}\int\sin(x)\,dx\Big|_{x=3t-1}=\frac{1}{3}(-\cos(x))\,\Big|_{x=3t-1}+C=-\frac{1}{3}\cos(3t-1)+C\,.$$

Case 2

The change of units is

$$z = h(y) = my + b\,.$$

Re-scaling the horizontal axis requires an adjustment of the integral:



Above, the y-axis is shrunk by a factor of 2, i.e., $z = y/2$. Therefore, the integral with respect to y will have multiplied by 2 to get one for z.

Theorem 2.4.6: Linear Composition Rule II

For any numbers m and b and any integrable function f, we have:

$$\boxed{\int \big(mf(x) + b\big)\, dx = m \int f\, dx + bx}$$

The result is just an instance of the Linearity Rule.

Example 2.4.7: space shift and flip

If y is the location and we change the place from which we start measuring, we have:

$$z = h(x) = y + y_0 \implies \int k\, dx = \int f\, dx + y_0 x\,.$$

If we change the direction of the x-axis, we have:

$$z = h(x) = -y \implies \int k\, dx = -\int f\, dx\,.$$

Example 2.4.8: miles to kilometers

Suppose x is the time and y is the location, then function h may represent the change of units of length, such as from miles to kilometers:

$$z = h(y) = 1.6y\,.$$

As we know, the quantities describing motion are simply replaced with their *multiples*. The new graphs are the vertically stretched versions of the old ones. Let's recast this statement in the integral form:

- If a is the acceleration with respect to miles, then the velocity with respect to kilometers is

$$\frac{1}{1.6} \int a\, dx\,.$$

- If v is the velocity with respect to miles, then the location with respect to kilometers is

$$\frac{1}{1.6}\int v\,dx\,.$$

Exercise 2.4.9

If a is the acceleration with respect to miles, then what is the position with respect to kilometers?

Exercise 2.4.10

Prove the above formulas.

Example 2.4.11: time and temperature

This is how *both cases* can appear. Recall the example from Chapter 2DC-4 when we have a function f that records the temperature (in Fahrenheit) as a function f of time (in minutes) replaced with another that records the temperature in Celsius as a function g of time in seconds:

- s time in seconds;
- m time in minutes;
- F temperature in Fahrenheit;
- C temperature in Celsius.

The conversion formulas are:

$$m = s/60\,,$$

and

$$C = (F-32)/1.8\,.$$

These are the relations between the four quantities:

$$g:\quad s \xrightarrow{\quad s/60\quad} m \xrightarrow{\quad f\quad} F \xrightarrow{\quad (F-32)/1.8\quad} C$$

And this is the new function:

$$F = k(s) = \big(f(s/60)-32\big)/1.8\,.$$

Then, we have:

$$\begin{aligned}\int k\,ds &= \int \big((f(s/60)-32)/1.8\big)\,ds\\ &= \int f(s/60)/1.8\,ds - \int 32/1.8\,ds\\ &= \frac{1}{1.8}\int f(s/60)\,ds - 32/1.8s\\ &= \frac{60}{1.8}\int f\,dm\bigg|_{m=s/60} - 32/1.8s\,,\end{aligned}$$

by the *Linear Composition Rule.*

Exercise 2.4.12

Provide a similar analysis for the sizes of shoes and clothing.

Example 2.4.13: degrees to radians

The conversion of the number of degrees y to the number of radians x is:

$$x = \frac{\pi}{180}y\,.$$

Then, for any function $z = f(x)$, we have:

$$\int f\left(\frac{\pi}{180}y\right)\, dy = \frac{180}{\pi}\int f\, dx\Bigg|_{x=\frac{\pi}{180}y}.$$

Because of the extra coefficient, the trigonometric integration formulas, such as

$$\int \sin x\, dx = -\cos x + C\,,$$

don't hold for degrees. Indeed, if we denote *sine and cosine for degrees* by $\sin_d y$ and $\cos_d y$ respectively, we have two entirely new functions:

$$\sin_d y = \sin\left(\frac{\pi}{180}y\right) \text{ and } \cos_d y = \cos\left(\frac{\pi}{180}y\right).$$

Therefore,

$$\begin{aligned}\int \sin_d y\, dy &= \frac{180}{\pi}\int \sin x\, dx\Bigg|_{x=\frac{\pi}{180}y}\\ &= \frac{180}{\pi}\cos x\Bigg|_{x=\frac{\pi}{180}y} + C\\ &= \frac{180}{\pi}\cos_d y + C\,.\end{aligned}$$

The formula just doesn't look as nice!

Example 2.4.14: logarithmic scale

The following *non-linear* change of units is called a *logarithmic scale*:

$$x = g(t) = 10^t\,.$$

Then, for a function $y = f(x)$, suppose F is its antiderivative. How do we, as we did above, express antiderivatives of $y = f(10^t)$ in terms of F? We would like to have a formula:

$$\int f(10^t)\, dt = ...$$

We proceed as before, by the Chain Rule:

$$\frac{d(F\circ g)}{dt} = \frac{dF}{dx}\Bigg|_{x=10^t}\cdot \left(10^t\right)' = f(10^t)10^t\ln 10\,.$$

Therefore,

$$F(10^t) = \int f(10^t)\cdot 10^t \ln 10\, dt\,,$$

and, further,

$$\int f(10^t)\cdot 10^t\, dt = \frac{1}{\ln 10}F(10^t)\,.$$

Unfortunately, the presence of the factor 10^t inside the integral seems to not allow us to finish the job and express *directly* the antiderivative of $y = f(10^t)$ in terms of F. We will need further analysis...

2.5. Integration by substitution: compositions

The linear substitution formula contains the derivative of the first, linear function:

$$\begin{aligned}\int f(mx+b)\,dx &= \frac{1}{m}\int f(u)\,du\Big|_{u=mx+b}\\ (mx+b)' &= m\end{aligned}$$

Because this derivative is just a number, it can be factored into the integral as it is moved to the other side:

$$\begin{aligned}(\sin(3x-1))' &= \cos(3x-1)\cdot(3x-1)'\\ &= \cos(3x-1)\cdot 3\\ \Longrightarrow \quad & \left(\frac{1}{3}\sin(3x-1)\right)' = \cos(3x-1)\\ \Longrightarrow \quad & \frac{1}{3}\sin(3x-1) = \int\cos(3x-1)\,dx\end{aligned}$$

So, our formula is, in truth, the following:

$$\boxed{\int f(mx+b)\cdot(mx+b)'\,dx = \int f(u)\,du\Big|_{u=mx+b}}$$

Now, how do we integrate functions with compositions and not just linear ones?

Just as with other integration formulas we, again, try to "reverse" the direction of differentiation.

Let's take $\sin(x^2)$. It is easy to differentiate by the *Chain Rule*:

$$\big(\sin(x^2)\big)' = \cos(x^2)\cdot 2x\,.$$

We would like to have a similar formula for the integral of this function:

$$\int \sin(x^2)\,dx = ?$$

But we don't recognize $\sin(x^2)$ as the derivative of any function we know... Can we see why? It's the extra factor g' we get every time we apply the Chain Rule to differentiate $f\circ g$.

Now, we *do* recognize $\cos(x^2)\cdot 2x$ from two lines above! Then,

$$\int \cos(x^2)\,2x\,dx = \sin(x^2) + C\,.$$

More examples? Here they are:

$$\int \sin(x^2)\,2x\,dx = -\cos(x^2)+C,\qquad \int e^{x^2}\,2x\,dx = e^{x^2}+C\,.$$

What do the three examples have in common? We see a pattern:

$$\begin{array}{lccccccc}\displaystyle\int & \cos & (x^2) & \cdot 2x & dx & = & \sin & (x^2)\\ \displaystyle\int & \sin & (x^2) & \cdot 2x & dx & = & -\cos & (x^2)\\ \displaystyle\int & e & ^{(x^2)} & \cdot 2x & dx & = & e & ^{(x^2)}\\ \displaystyle\int & ? & (x^2) & \cdot 2x & dx & = & ? & (x^2)\end{array}$$

Everything is the same except whatever is behind these question marks.

We know what is missing and we rewrite:

$$\boxed{\int f(x^2)\cdot 2x\,dx = F(x^2)+C \quad \text{with } F' = f}$$

In other words, F is an antiderivative of f.

So, to integrate these, we need to solve this problem:

- ▶ Given f, find F with $F' = f$.

This is, of course, also an *integration problem*, but not with respect to x! What is this variable? Let's decompose:

$$x \mapsto x^2 = u \mapsto f(u) = z$$

So, both f and F are functions of some u, an intermediate variable, that we've made up. Then, to find F, we integrate f with respect to u:

$$\boxed{F(u) = \int f(u)\,du}$$

This is a *change of variables*!

Example 2.5.1: composition with x^2

Evaluate:

$$\int \underbrace{\sqrt[3]{x^2}}_{\text{decompose}} \cdot 2x\,dx = \,?$$

The key step is to break the composition apart, to find u, f, F. So, $u = x^2$, $f(u) = u^{1/3}$. Then,

$$F(u) = \int \sqrt[3]{u}\,du = \int u^{\frac{1}{3}}\,du \overset{\text{PF}}{=\!=} \frac{u^{\frac{1}{3}+1}}{\frac{1}{3}+1} + C = \frac{3}{4}u^{\frac{4}{3}} + C\,.$$

Even though integration is finished, this isn't the answer because it has to be a function of x! We need to substitute $u = x^2$ back into this function:

$$F(x^2) = \frac{3}{4}\left(x^2\right)^{\frac{4}{3}} + C\,.$$

For a more general analysis, we replace x^2 with $g(x)$. We are prepared to integrate this (and nothing else):

$$\int f(g(x))\cdot g'(x)\,dx\,.$$

The answer is $F(g(x))$, where F is an antiderivative of f:

$$F' = f\,.$$

Theorem 2.5.2: Integration by Substitution

Given a continuous function f and a differentiable function g, we have:

$$\boxed{\int f(g(x))\cdot g'(x)\,dx = F(g(x)) + C}$$

where F is any antiderivative of f:

$$\boxed{F(u) = \int f(u)\, du}$$

Proof.

$$\begin{aligned}(F(g(x)))' &\overset{\text{CR}}{=\!=} F'(g(x) \cdot g'(x) \\ &= f(g(x))g'(x)\,.\end{aligned}$$

Conclusion: we can integrate compositions when the following is satisfied:

- The derivative of the "inside" function is present as a factor.

In other words, as a *prerequisite*, we need to have the following:

$$\int f(\underbrace{g(x)}_{\text{"inside" function}}) \cdot \underbrace{g'(x)}_{\text{its derivative}}\, dx$$

Example 2.5.3: application of theorem

Evaluate

$$\int \sqrt{x^3+1} \cdot 3x^2\, dx\,.$$

Observe first that the derivative of the function inside is present:

$$(x^3+1)' = 3x^2\,.$$

So, the theorem should work:

$$\begin{array}{ll} \text{decomposition:} & \text{integration:} \\ f(u) = \sqrt{u} & \Longrightarrow F(u) = \int u^{\frac{1}{2}}\, du = \frac{2}{3}u^{\frac{3}{2}} + C \\ u = g(x) = x^3+1 & \Longrightarrow g'(x) = 3x^2 \\ \text{back-substitution:} & F(g(x)) = \frac{2}{3}(x^3+1)^{\frac{3}{2}} + C \end{array}$$

Note how we converted the original integral to a simpler one, with respect to u.

Example 2.5.4: integration by substitution

Evaluate

$$\int \sqrt{x^3+1} \cdot x^2\, dx = ?$$

Just notice that $(x^3+1)' = 3x^2$, not x^2. The condition doesn't seem to be satisfied anymore... However, this function is just a constant multiple of the one in the last example and so, therefore, is the integral. We will ignore this shortcut though.

We'll try to apply the *Integration by Substitution* formula anyway:

$$\left.\begin{array}{lll} \text{first substitution:} & u & = x^3+1 \\ \text{second substitution:} & u' & = 3x^2 \end{array}\right\} \begin{array}{l} \text{we convert the integral with respect to } x \\ \text{to a new one with respect to } u \end{array}$$

The hope is that the new integral will be simpler than the original.

We already have all we need here. We break what's inside the integral apart but not in the obvious way:

$$\sqrt{x^3+1}=\sqrt{u}, \quad x^2=\frac{1}{3}u' .$$

Now we deal with the integral itself.

$$\begin{aligned}
\int \underbrace{\sqrt{x^3+1}}_{u \text{ inside}}\, x^2\,dx &= \int \sqrt{u}\cdot\frac{1}{3}\,du \\
&= \frac{1}{3}\int u^{\frac{1}{2}}\,du && \text{A new integral.}\\
&\overset{\text{PF}}{=\!=} \frac{1}{3}\frac{2}{3}u^{\frac{3}{2}}+C && \text{Integration finished.}\\
&= \frac{1}{3}\frac{2}{3}(x^3+1)^{\frac{3}{2}}+C && \text{Back-substitution } u=x^3+1 .
\end{aligned}$$

Answer:

$$\int \sqrt{x^3+1}\,x^2\,dx=\frac{2}{9}(x^3+1)^{\frac{3}{2}}+C .$$

Exercise 2.5.5

Evaluate

$$\int \sqrt{x^3+1}\,x^2\,dx .$$

Exercise 2.5.6

Evaluate

$$\int \sqrt{x^4+1}\,x^3\,dx .$$

We can re-write our theorem as a single formula as follows:

$$\boxed{\int f(g(x))\cdot g'(x)\,dx=\int f(u)\,du\Big|_{u=g(x)}}$$

In this version, there is no confusion about whether the integration in the right-hand side has been carried out.

The simplest kind of integral for which this approach *always* works involves a linear change of variables. If

$$g(x)=mx+b,\ m\neq 0 ,$$

our formula becomes:

$$\int f(mx+b)\cdot m\,dx=\int f(u)\,du\Big|_{u=mx+b} .$$

We consequently recover the familiar *Linear Composition Rule*:

$$\int f(mx+b)\,dx=\frac{1}{m}\int f(u)\,du\Big|_{u=mx+b} .$$

Example 2.5.7: one-line integration

Evaluate:

$$\int e^{3x}\,dx = \frac{1}{3}\int e^u\,du\Bigg|_{u=3x} = \frac{1}{3}e^u + C\Bigg|_{u=3x} = \frac{1}{3}e^{3x} + C\,.$$

Example 2.5.8: change of variables

Evaluate:

$$\int e^x \sin(e^x)\,dx\,.$$

Let's break down the composition, $\sin(e^x)$:

$$u = e^x, \quad y = \sin u\,.$$

Furthermore,

$$u' = e^x.$$

Use these two:

$$\begin{aligned} \int e^x \sin(e^x)\,dx &= \int \sin u\,du && \text{Evaluate.} \\ &= -\cos u + C && \text{Substitute.} \\ &= \cos e^x + C\,. \end{aligned}$$

Exercise 2.5.9

Evaluate:

$$\int \sqrt{\sin x}\cdot \cos x\,dx\,.$$

Exercise 2.5.10

Evaluate:

$$\int e^{e^x + x}\,dx\,.$$

Example 2.5.11: no composition

$$\begin{aligned} \int \tan x\,dx &= && \text{What, no composition?!} \\ &= \int \frac{\sin x}{\cos x}\,dx && \text{There is a division though.} \\ &= \int \sin x \cdot \frac{1}{\cos x}\,dx && \text{It's a multiplication, in fact.} \\ &= \int \sin x\,(\cos x)^{-1}\,dx && \text{So, there is a composition after all!} \\ &= -\int (\cos x)'\,(\cos x)^{-1}\,dx && \text{And the derivative of the inside function is present.} \\ &= -\int (u)^{-1}\,du && \text{The formula applies with } u = \cos x. \\ &= -\ln u + C && \text{We integrate.} \\ &= -\ln \cos x + C\,. && \text{We back-substitute.} \end{aligned}$$

Warning!

We make sure that the integrals that we face *can* be evaluated with this method.

2.6. Change of variables in integrals

Example 2.6.1: failure after substitution

Let's take another look at this integral:

$$\int xe^{x^2}\,dx = \frac{1}{2}\int e^u\,du = e^u + C = e^{x^2} + C\,.$$

It works so well! Changing the power, x to x^2, ruins this nice arrangement:

$$\int x^2e^{x^2}\,dx = \int ue^u\,dx = \text{... now what?}$$

In fact, no power of x other than 1 will allow integration by substitution according to the formula:

$$\int x^3e^{x^2}\,dx = ? \quad \int x^4e^{x^2}\,dx = ? \quad \int x^{1/2}e^{x^2}\,dx = ? \quad \int x^{\pi}e^{x^2}\,dx = ?$$

Warning!

Do not replace x in dx with u, $dx \neq du$!

We still would like to be able to *convert* an integral to a new variable. It is always possible! The challenge is what to do with dx.

We have to look at the integral differently. *What* exactly do we integrate? In the integral,

$$\int k(x)\,dx\,,$$

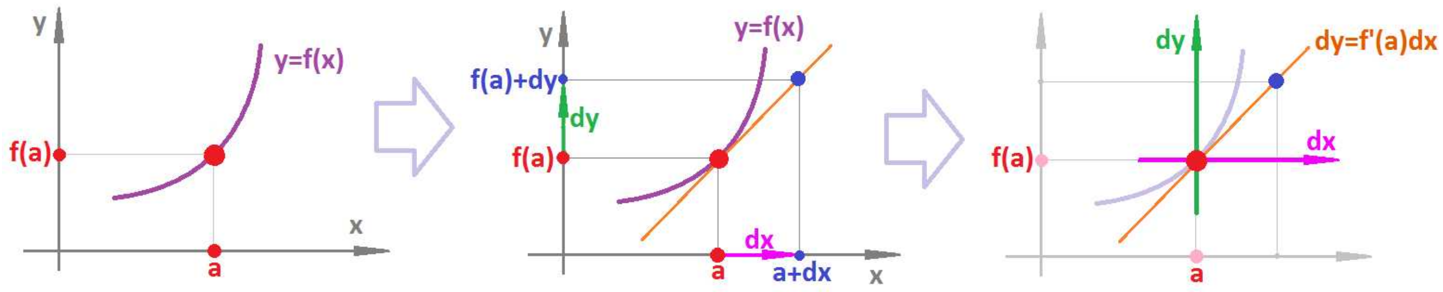
it has been a function, $k(x)$, while $\int$ and dx serve as mere brackets. This doesn't work anymore! We need to make sense of dx. Significantly, we switch from a function to a *differential form*, $k(x)\cdot dx$. As presented in Chapter 2DC-4, the form comes from the following:

$$y = f(x) \text{ at } x = a \implies \frac{dy}{dx} = f'(a)\,,$$

and, furthermore,

$$\implies \quad dy = f'(a)\cdot dx\,.$$

This is a relation between the two *extra variables*, nothing but numbers, once the relation between the old ones has been specified:



In the graph, dx is the run and dy is the rise of the tangent line. They are called the *differentials* of x and y respectively. The dependence between them varies from location to location. This is further discussed in Chapter 4.

Back to integration.

So, dx is the differential of x, which is a variable separate from, but related to, x. Then, $f'(x) \cdot dx$ is just a function of two variables. The dependence of the differential form on the second variable is especially simple; it's a multiple.

Now integration by substitution.

The differential du of u is a variable separate from, but related to, u. So, changing variables means going from x to u *and* from dx to du. Handling the differentials is a separate step in the process of changing the variable.

Recall how the *Chain Rule*, in the Leibniz notation, is interpreted as, and it is, a "cancellation" of du (when it's not zero):

$$\boxed{\frac{dy}{dx} = \frac{dy}{\cancel{d}u}\frac{\cancel{d}u}{dx}}$$

A similar idea applies to integrals. We take the formula of integration by substitution and switch to the Leibniz notation, subject to the substitution:

$$\int f(g(x)) \cdot g'(x)\, dx = \int f(u)\, du \Bigg|_{u=g(x)}$$

and

$$\int f(u) \cdot \frac{du}{\cancel{d}x}\ \cancel{d}x = \int f(u)\, du\,.$$

Note how dx "cancels", turning the integral with respect to x to one with respect to u. We take this idea one step further.

Corollary 2.6.2: Change of Variables of Differential Form

Under a substitution $u = g(x)$ in an integral, we also substitute:

$$\boxed{du = g'\, dx}$$

This formula is used – in addition to $u = g(x)$ – in order to complete the substitution.

Example 2.6.3: change of variables is the goal

Let's apply the formula to the example from the last section:

$$\int e^x \sin(e^x)\, dx\,.$$

We start with a substitution this time:

$$u = e^x \implies du = e^x dx \implies dx = \frac{du}{e^x}$$

Substitute *both* the first and the last of these into the integral:

$$\begin{aligned}\int e^x \sin(e^x)\, dx &= \int e^x \sin e^x \frac{du}{e^x} && \text{We substitute the last and cancel.}\\ &= \int \sin u\, du && \text{We substitute the first, and the change of variables is complete!}\\ &= -\cos u + C && \text{The rest is a bonus.}\\ &= -\cos e^x + C\,.\end{aligned}$$

With the formula, we can change variables in any integral, even the kind that's not subject to integration by substitution.

Example 2.6.4: change of variables only

Let's evaluate:

$$\int x^2 e^{x^2}\, dx = \,?$$

The change of variables is the same as before ($x \geq 0$):

$$u = x^2 \implies du = 2x\, dx \implies dx = \frac{du}{2x}$$

However, anticipating that the cancellation might not be as easy as last time, we also find the *inverse substitution* (literally the inverse of the substitution function):

$$u = x^2 \implies x = \sqrt{u}\,.$$

We have *three* substitutions:

1. $x^2 = u$
2. $dx = \dfrac{du}{2x}$
3. $x = \sqrt{u}$

Substitute:

$$\begin{aligned}\int x^2 e^{x^2}\, dx &= \int x^2 e^u\, dx && \#1\\ &= \int x^2 e^u \frac{du}{2x} && \#2\\ &= \int \left(\sqrt{u}\right)^2 e^u \frac{du}{2\sqrt{u}} && \#3\\ &= \frac{1}{2}\int \sqrt{u} e^u\, du\,.\end{aligned}$$

Even though the change of variables hasn't made the integral easier to integrate, the conversion is complete!

We can have *any* substitution in *any* integral.

Example 2.6.5: bad substitution

Let's pick a *wrong* change of variables in the familiar integral:

$$\int x e^{x^2}\, dx = \,?$$

The following substitution is chosen even though we don't anticipate that it will simplify the integral:

$$\#1.\ u = x^3\,.$$

The differential is found:

$$\#2.\ u = x^3 \implies du = 3x^2\,dx \implies dx = \frac{du}{3x^2}\,.$$

The inverse substitution is found:

$$\#3.\ x = u^{1/3}\,.$$

Substitute:

$$\begin{aligned}\int xe^{x^2}\,dx &= \int xe^{(u^{1/3})^2}\,dx && \#3\\ &= \int xe^{u^{2/3}}\,\frac{du}{3x^2} && \#2\\ &= \int u^{1/3}e^{u^{2/3}}\,\frac{du}{3\left(u^{1/3}\right)^2} && \#3\\ &= \frac{1}{3}\int u^{-1/3}e^{u^{2/3}}\,du\,.\end{aligned}$$

Our choice of a new variable was unwise.

Exercise 2.6.6

Carry out the substitution $u = x^4$ in the above integral.

Exercise 2.6.7

Carry out the substitution $u = x$ in the above integral.

Exercise 2.6.8

Make up your own integral and carry out the substitution $u = x^2$. Repeat.

Example 2.6.9: clue for new variable

If we hope to simplify the integral by substitution, the new variable should be equal to the "inside" function of the composition. For example, consider:

$$\int \sqrt{x+1}\cdot x\,dx\,.$$

We choose $u = x + 1$. Then, $du = dx$. Therefore,

$$\begin{aligned}\int \sqrt{x+1}\cdot x\,dx &= \int u^{1/2}(u-1)\,du && \text{Is it any better?}\\ &= \int u^{1/2}u\,du + \int u^{1/2}(-1)\,du && \text{Yes!}\\ &= \int u^{3/2}u\,du - \int u^{1/2}\,du\\ &= \frac{2}{5}u^{5/2} - \frac{2}{3}u^{3/2} + C\\ &= \frac{2}{5}(x+1)^{5/2} - \frac{2}{3}(x+1)^{3/2} + C\,.\end{aligned}$$

Example 2.6.10: logarithmic change of variables

Let's revisit the issue of converting units to a *logarithmic scale*:

$$x = 10^t\,.$$

Then,

$$dx = 10^t \ln 10\, dt \implies dt = \frac{dx}{10^t \ln 10}\,.$$

Substitute into the integral and simplify:

$$\int f(10^t)\, dt = \int f(x)\, \frac{dx}{10^t \ln 10} = \frac{1}{\ln 10} \int f(x) \frac{1}{x}\, dx\,.$$

We have expressed the integral of $y = f(10^t)$ as an integral with respect to x.

This is the *summary* of what is going on. We start with a familiar diagram for the Chain Rule of differentiation:

$$\begin{tikzcd} F(g(x)) \arrow[r, "\frac{d}{dx}\square"] \arrow[d, "u=g(x)"', "\text{substitution}"'] & F'(g(x))g'(x) \\ F(u) \arrow[r, "\frac{d}{du}\square"] & F'(u) \arrow[u, "\text{CR with } u=g(x)"'] \end{tikzcd}$$

Now we remake the diagram into one about integration by reversing the horizontal arrows:

$$\begin{tikzcd} F(g(x)) \arrow[d, "u=g(x)", "\text{substitution}"'] & F'(g(x))g'(x) \arrow[l, "\int \square\, dx"'] \\ F(u) & F'(u) \arrow[l, "\int \square\, du"'] \arrow[u, "\text{CR with } u=g(x)"'] \end{tikzcd}$$

We re-name the function for convenience:

$$\begin{tikzcd} F(g(x)) & f(g(x))g'(x) \arrow[l, "\int \square\, dx"'] \\ F(u) \arrow[u, "u=g(x)"', "\text{substitution}"] & f(u) \arrow[l, "\int \square\, du"'] \arrow[u, "\text{CR with } u=g(x)"'] \end{tikzcd}$$

Thus the change of variables method of integration gives us an alternative way of getting from top right to top left (integration with respect to x). We take a detour by following the clockwise path around the square:

1. the Chain Rule formula in reverse
2. integration with respect to u
3. back-substitution

Example 2.6.11: substitution with diagram

The computation of the integral

$$\int \cos(x^2)\, 2x\, dx = \sin(x^2) + C$$

is illustrated below:

$$\begin{array}{cccc} & \sin(x^2) & \xleftarrow{\int \square\, dx = \ ?} & \boxed{\cos(x^2)\,2x} \\ \text{substitution} & \uparrow{\scriptstyle u=x^2} & & \uparrow{\scriptstyle \text{CR with } u=x^2} \\ & \sin(u) & \xleftarrow{\int \square\, du} & \cos(u) \end{array}$$

Exercise 2.6.12

Execute the substitution $u = e^x$ for the integral (don't evaluate the resulting integral):

$$\int \sin(1+e^x)\,dx\,.$$

2.7. Change of variables in definite integrals

At the next stage, we arrive at the following question:

▶ What happens when we use substitution to evaluate *definite* integrals?

First, nothing *has to* change. After all, according to the *Fundamental Theorem of Calculus*, all we need is an antiderivative. So, to find the definite integral

$$\int_a^b f(g(x))g'(x)\,dx\,,$$

find the corresponding indefinite integral first, as we did in the last section,

$$H(x) = \int f(g(x))g'(x)\,dx\,,$$

if possible. If it is, then the last step is as *simple* as it gets:

$$\int_a^b f(g(x))g'(x)\,dx = H(b) - H(a)\,.$$

Example 2.7.1: using FTC for old variable

Let's evaluate:

$$\int_0^1 e^x \sin(e^x)\,dx\,.$$

We have already found the antiderivative in the last section:

$$\int e^x \sin(e^x)\,dx = -\cos e^x + C\,.$$

Therefore,

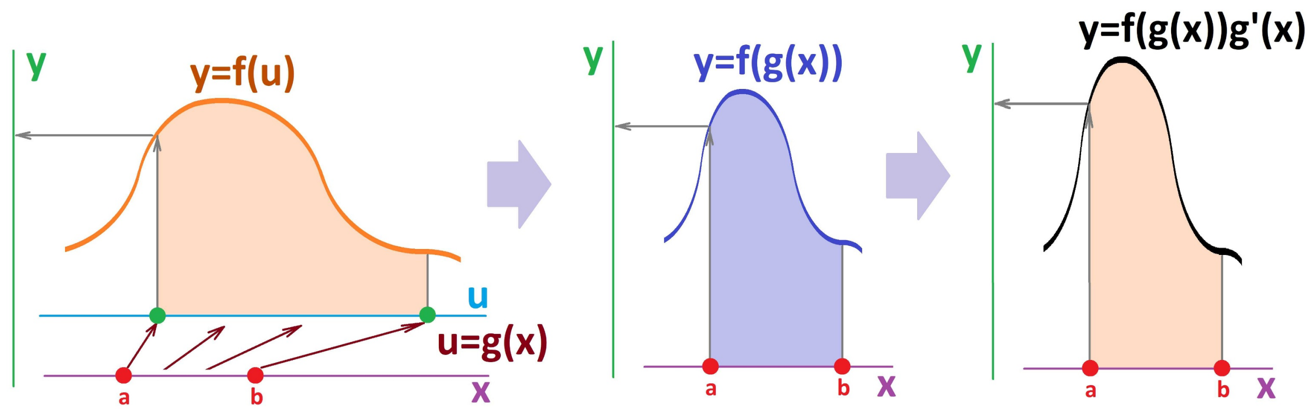
$$\int_0^1 e^x \sin(e^x)\,dx = -\cos e^1 - (-\cos e^0) = -\cos e + \cos 1\,.$$

Done!

A better question is:

▶ What happens to the definite integral under a *change of variables*, in addition to the substitutions presented in the last section?

The substitution is just a transformation of the x-axis. It, therefore, shrinks/stretches the *domain of integration*:



So, we have to track, the end-points of the domain, i.e., the bounds of the integral, under our transformation.

Example 2.7.2: using FTC for new variable

Let's take a closer look at the computation in the last example:

$$\begin{aligned}\int e^x \sin(e^x)\,dx &= -\cos u + C && = -\cos e^x + C\\ \int_0^1 e^x \sin(e^x)\,dx &= -\cos e^1 - (-\cos e^0) && = -\cos e + \cos 1\end{aligned}$$

It is done via the substitution $u = e^x$. We realize that we could have jumped from $-\cos u + C$ to $-\cos e + \cos 1$ by omitting the back-substitution step in the upper right corner! Indeed:

$$\cos u \Big|_1^e = -\cos e + \cos 1\,.$$

We just need to see the relation between the *bounds* of the two integrals, with respect to x and u:

$$\begin{aligned} x = 0 \ &\mapsto\ u = e^0 = 1\\ x = 1 \ &\mapsto\ u = e^1 = e\end{aligned}$$

The back-substitution becomes redundant.

So, under the change of variable $u = g(x)$, the domain of integration changes from

- $[a, b]$ for x to
- $[g(a), g(b)]$ for u.

Even thought the stretch/shrink might be non-uniform, we only care about the end-points.

Example 2.7.3: no back-substitution

Find the area under the graph of $y = x^2 \cos x^3$ from 0 to 2. We have:

$$\text{Area} = \int_0^2 x^2 \cos x^3\,dx\,.$$

Substitution first:

$$u = x^3 \implies du = 3x^2\,dx \implies dx = \frac{du}{3x^2}\,.$$

Then,

$$\int x^2 \cos x^3 \, dx = \int x^2 \cos u \, \frac{du}{3x^2} = \frac{1}{3} \int \cos u \, du \, .$$

Now, what would this computation look like for the definite integral? Let's make it clear what variables we are referring to:

$$\int_{x=0}^{x=2} x^2 \cos x^3 \, dx = \int_{x=0}^{x=2} x^2 \cos u \, \frac{du}{3x^2} = \frac{1}{3} \int_{x=0}^{x=2} \cos u \, du \, .$$

We have mismatched variables!

In order to fix that, we find the domain of integration by finding the bounds for u from the corresponding bounds for x:

$$\begin{aligned} x = 0 \ &\mapsto \ \ u = 0^3 = 0 \\ x = 2 \ &\mapsto \ \ u = 2^3 = 8 \end{aligned}$$

So, $[0, 2]$ for x becomes $[0, 8]$ for u. Then,

$$\int_{x=0}^{x=2} x^2 \cos x^3 \, dx = \frac{1}{3} \int_{x=0}^{x=2} \cos u \, du = \frac{1}{3} \int_{u=0}^{u=8} \cos u \, du \, .$$

The change of variables in the definite integral is complete.

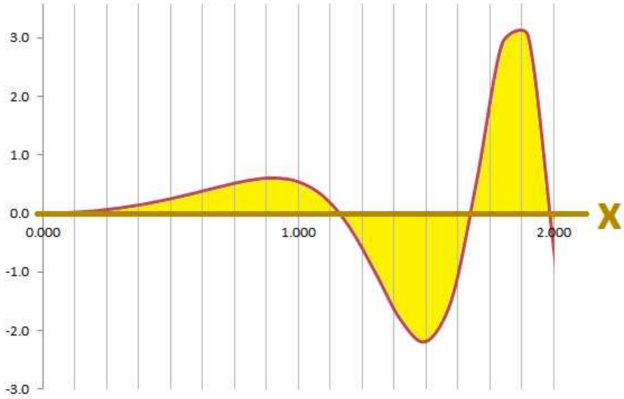
We don't have to go back to x in order to finish the computation:

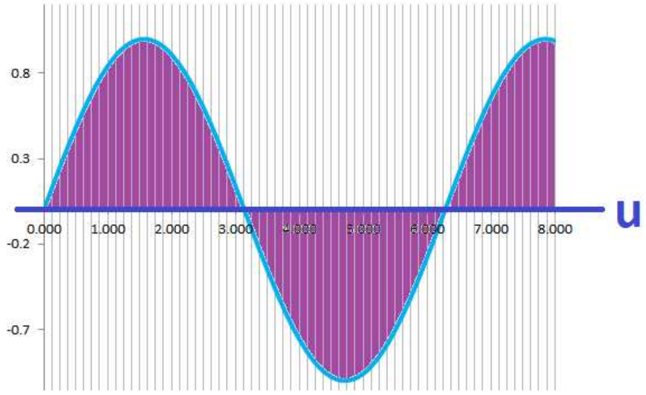
$$\text{Area } = \frac{1}{3} \int_{u=0}^{u=8} \cos u \, du \ \overset{\text{FTC}}{=\!=\!=} \ \sin u \Bigg|_{u=0}^{u=8} = \sin 8 - \sin 0 = \sin 8 \, .$$

In summary, we showed that the areas of these two regions under these graphs,

$$y = x^2 \cos x^3 \ \text{ and } \ y = \frac{1}{3} \cos u \, ,$$

are equal and then found the latter one:





This is the summary:

Corollary 2.7.4: Definite Integration by Substitution

Under a substitution $u = g(x)$ in a definite integral, we have:

$$\int_a^b f(g(x)) \cdot g'(x) \, dx = F(u) \Bigg|_{g(a)}^{g(b)} = F(g(b)) - F(g(a)) \, ,$$

where F is any antiderivative of f.

Proof.

Recall the formula of *integration by substitution*:

$$\int f(g(x)) \cdot g'(x)\, dx = \int f(u)\, du \bigg|_{u=g(x)} = F(g(x))\,.$$

The formula in the theorem follows now from the *Fundamental Theorem of Calculus.*

Example 2.7.5: no back-substitution

Let's carry out a change of variables in a definite integral *without back-substitution.* Consider

$$\int_0^\pi \cos^2 x \sin x\, dx\,.$$

The initial part – choosing a substitution – remains the same. We notice a composition and choose the function on the inside to be the substitution function for the new variable:

$$\cos^2 x = (\cos x)^2 \implies u = \cos x\,.$$

The second step is to find the rest of the substitution:

$$u = \cos x \implies du = -\sin x\, dx \implies dx = \frac{du}{-\sin x}$$

The next step is to find the corresponding bounds for u:

$$\begin{aligned} x = \pi &\implies u = \cos\pi = -1 \\ x = 0 &\implies u = \cos 0 = 1 \end{aligned}$$

We convert the integral to u now and then evaluate it:

$$\begin{aligned} \int_0^\pi \cos^2 x \sin x\, dx &= \int_0^\pi (\cos x)^2 \sin x\, dx \\ &= \int_1^{-1} (u)^2 \sin x \, \frac{du}{-\sin x} \\ &= \int_1^{-1} (u)^2\, \frac{du}{-1} \\ &= -\int_1^{-1} u^2\, du \\ &= \int_{-1}^{1} u^2\, du \\ &= \frac{1}{3}u^3 \bigg|_{u=-1}^{u=1} \\ &= \frac{1}{3}\big(1^3 - (-1)^2\big) \\ &= \frac{2}{3}\,. \end{aligned}$$

This is what happens to integration by substitution when we proceed to definite integration. The extra step

is the Fundamental Theorem of Calculus, for either variable:

$$\begin{array}{ccccccc} & f(g(x))g'(x) & \xrightarrow{\int dx} & F(g(x)) & \xrightarrow{\text{FTC}} & I & \\ \text{substitution} & \uparrow \text{CR} & & \downarrow u=g(x) & & || & \text{same!} \\ & f(u) & \xrightarrow{\int du} & F(u) & \xrightarrow{\text{FTC}} & I & \end{array}$$

Thus, the result of definite integration – a number – is the same no matter what variable we choose.

Here's a more explicit way to write our formula:

$$\boxed{\int_{x=a}^{x=b} f(g(x)) \cdot g'(x)\, dx = \int_{u=g(a)}^{u=g(b)} f(u)\, du}$$

This is the formula of change of variables in definite integral written in the Leibniz notation:

$$\boxed{\int_a^b f(u) \cdot \frac{du}{dx}\, dx = \int_{u(a)}^{u(b)} f(u)\, du}$$

under a substitution $u = u(x)$.

Exercise 2.7.6

Execute the substitution $u = e^x$ for the integral (don't evaluate the resulting integral):

$$\int_1^2 \cos(1 - e^x)\, dx\,.$$

2.8. Trigonometric substitutions

Back to indefinite integrals...

Which substitution to choose might not be always obvious, and when it is, it might lead to an integral that isn't any simpler that the original. The latter problem is especially common. Consider the familiar integral:

$$\begin{array}{ccccc} ??? & & ??? & & \frac{1}{2}\int e^u\, du \\ & \nwarrow_{u=e^x} & \uparrow & \nearrow^{u=x^2} & \\ ??? & \xleftarrow{u=\sin x} & \boxed{\int xe^{x^2}\, dx} & \xrightarrow{u=x^3} & \frac{1}{3}\int u^{-1/3}e^{u^{2/3}}\, du \\ & \swarrow^{u=\cos x} & \downarrow & \searrow^{u=x^4} & \\ ??? & & ??? & & \frac{1}{4}\int u^{-1/2}e^{u^{1/2}}\, du \end{array}$$

Even though the change of variables has been carried out correctly, it might still be a dead end!

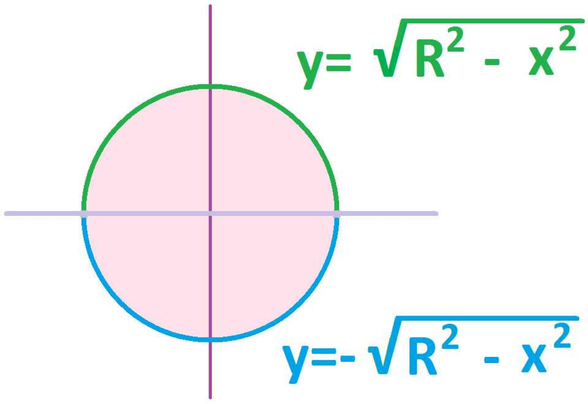
Exercise 2.8.1

Carry out these substitutions.

Sometimes we have to come up with entirely new ideas...

Let's revisit the question addressed in the last chapter, *what is the area of a circle* of radius R?

$$\text{Area } = 2\int_{-R}^{R}\sqrt{R^2-x^2}\,dx = \pi R^2\,.$$



To prove the formula, we need to integrate this:

$$\int\sqrt{R^2-x^2}\,dx\,.$$

How?

There is a composition... Let's try substitution! The obvious choice is:

$$u = R^2-x^2 \implies du = -2x\,dx\,.$$

Substitute:

$$\int\sqrt{R^2-x^2}\,dx = \int\sqrt{u}\,\frac{du}{-2x} = \int\sqrt{u}\,\frac{du}{-2\sqrt{R^2-u}} = -\frac{1}{2}\int\sqrt{\frac{u}{R^2-u}}\,du\,.$$

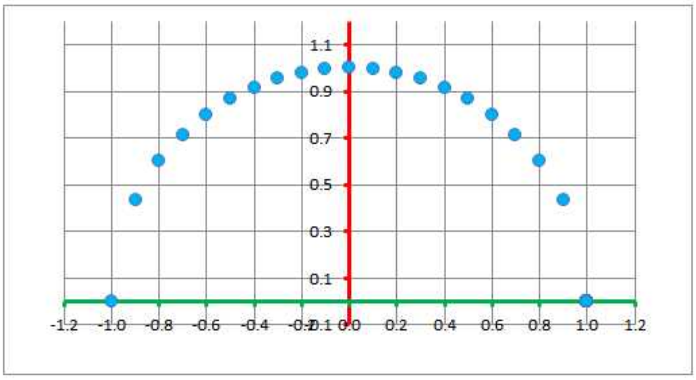
Change of variables is completed but, unfortunately, the new integral is no simpler than the original!

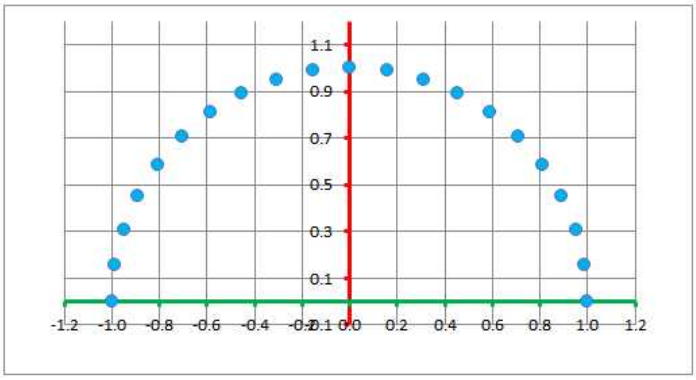
Let's take another look at the integrand:

$$y = \sqrt{R^2-x^2}\,.$$

What is its graph? It is upper half of the circle of radius R given directly (explicitly) as the graph of this function. The circle is also given by a *relation* (implicitly) by $x^2+y^2=R^2$. There may be a third possibility, if we just find a better variable...

On the left, we interpret the graph representation of the circle visualized as *motion*: time on the x-axis and location on the y-axis. As a result, the dots appear at equal intervals of time, i.e., horizontally:

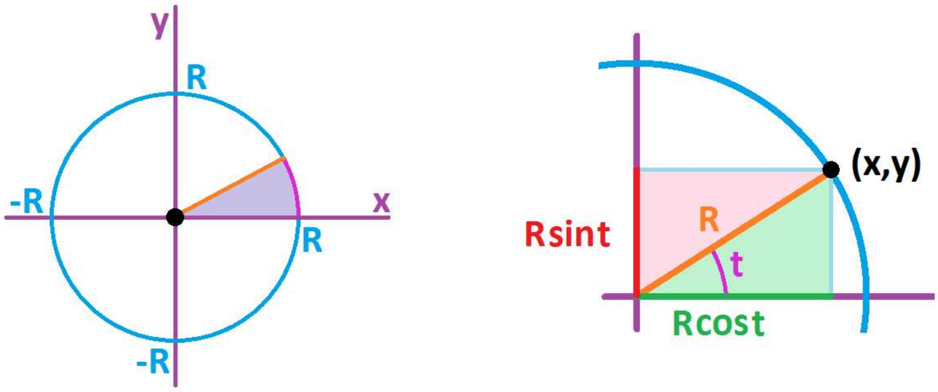




What we can see is how motion starts fast, then slows down to almost zero in the middle, and then accelerates again. But what if we consider instead a simple *rotation* on the plane? Such a rotation would

progress through the angles at a constant rate, shown on right. So, maybe the *angle*, say t, should be our new variable? Then, the formulas for x and y come from the basic trigonometry:

$$\begin{cases} x = R\cos t\,, \\ y = R\sin t\,. \end{cases}$$



The rotation is counterclockwise starting from $(R,0)$ and t runs from 0 to π.

A new variable has appeared naturally:

$$\boxed{x = R\cos t}$$

This is our substitution.

The main difference from previous examples of integration by substitution is that instead of the old variable given in terms of the new, as in:

$$u = x^2\,,$$

here, the new variable is given in terms of the old. This is why such a formula is often called an *inverse* substitution. No matter! We have:

$$t = \cos^{-1}(x/R)\,,$$

with

$$-\pi/2 \le t \le \pi/2\,.$$

In fact, to carry out this change of variables, all we need is this:

$$dx = -R\sin t\,dt\,.$$

We substitute but, in order to simplify, we'll also need the *Pythagorean Theorem* (Volume 1):

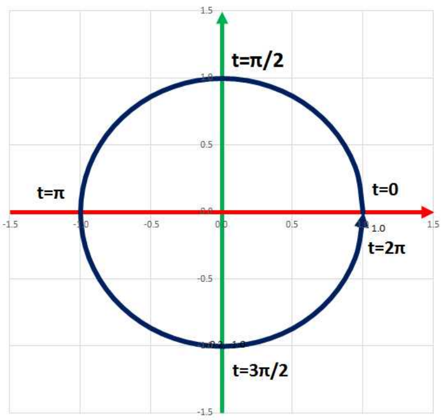
$$\sin^2 t + \cos^2 t = 1\,.$$

Then, we find an antiderivative with respect to t as follows:

$$\begin{aligned}
\int \sqrt{R^2 - x^2}\,dx &= \int \sqrt{R^2 - \cos^2 t}\cdot(-R\sin t\,dt) && \text{Use PT.}\\
&= \int R\sin t\cdot(-R\sin t\,dt) && \text{Change of variables finished.}\\
&= -R^2\int \sin^2 t\,dt && \text{A trig formula next.}\\
&= -R^2\int \frac{1-\cos 2t}{2}\,dt\\
&= -\frac{R^2}{2}\left(\int dt - \int \cos 2t\,dt\right)\\
&= -\frac{R^2}{2}\left(t - \frac{1}{2}\sin 2t\right)\,.
\end{aligned}$$

Integration is finished.

We won't do back-substitution because our interest is a definite integral. We only need to know the bounds for the new variable.



We find them from the picture above. Or from algebra:

$$x = -R \implies t = \pi \quad \text{and} \quad x = R \implies t = 0$$

Substitute:

$$\begin{aligned} \text{Area of a circle of radius } R &= 2 \cdot \int_{x=-R}^{x=R} \sqrt{R^2 - x^2}\, dx \\ &= -2 \cdot \frac{R^2}{2} \left(t - \frac{1}{2} \sin 2t \right) \Bigg|_{t=\pi}^{t=0} \\ &= -R^2 \left(0 - \frac{1}{2} \sin(2 \cdot 0) - (\pi + \frac{1}{2} \sin 2\pi) \right) \\ &= \pi R^2 \,. \end{aligned}$$

Exercise 2.8.2

Modify the proof for the substitution $x = R \sin t$.

Example 2.8.3: direct or inverse

Recall how we evaluated the integral:

$$\int x^2 e^{x^2}\, dx = \,?$$

The substitution was chosen to be $u = x^2$ from which we derived the other two items:

$$\begin{array}{lll} 1. & x^2 & = u \\ 2. & dx & = \dfrac{du}{2x} \\ 3. & x & = \sqrt{u} \end{array} \qquad \longrightarrow \qquad \begin{array}{lll} 1. & x & = \sqrt{u} \\ 2. & dx & = \dfrac{du}{2\sqrt{u}} \\ 3. & x^2 & = u \end{array}$$

However, we could have started with #3 (the inverse) with the same result!

The strategy is summarized below with two extra options:

Trigonometric Substitutions

Suppose $a > 0$. Then:

- When the integrand contains $\sqrt{a^2 - x^2}$, or sometimes $a^2 - x^2$, use the substitution:

$$x = a \sin t \quad \text{or} \quad x = a \cos t \,.$$

- When the integrand contains $\sqrt{a^2+x^2}$, or sometimes a^2+x^2, use the substitution:

$$x = a\tan t\,.$$

- When the integrand contains $\sqrt{x^2-a^2}$ (note the sign), or sometimes x^2-a^2, use the substitution:

$$x = a\sec t\,.$$

Warning!

This is not a theorem and it doesn't guarantee success.

Example 2.8.4: simplification

Let's simplify this:

$$\int (4-x^2)^{3/2}\,dx\,.$$

The expression matches option (1):

$$\begin{aligned} & a^2 && - && x^2 \\ & 4 && - && x^2 \\ \Longrightarrow\ & a=2 \end{aligned}$$

Therefore, we try:

$$x = 2\sin t\,.$$

Then,

$$dx = 2\cos t\,dt\,.$$

This is our key simplification:

$$4-x^2 = 4-(2\sin t)^2 = 4-4\sin^2 t = 4(1-\sin^2 t) = 4\cos^2 t\,.$$

Substitute:

$$\int (4-x^3)^{3/2}\,dx = \int (4\cos^2 t)^{3/2}(2\cos t\,dt) = 16\int \cos^4 t\,dt\,.$$

Exercise 2.8.5

Evaluate this integral:

$$\int \frac{1}{1+x^2}\,dx\,.$$

Exercise 2.8.6

Evaluate this integral:

$$\int \sqrt{x^2-1}\,dx\,.$$

2.9. Integration by parts: products

The *Product Rule for Derivatives* expresses the derivative of the product of two functions in terms of their derivatives (and the functions themselves):

$$(f \cdot g)' = f' \cdot g + f \cdot g' .$$

There is no "Product Rule for integrals" that would express the integral of the product of two functions in terms of their integrals (and the functions themselves):

$$\int (f \cdot g)\, dx = ?$$

Let's nonetheless try to get whatever we can from PR. We integrate it:

$$\begin{aligned} \int (f \cdot g)'\, dx &= \int (f' \cdot g + f \cdot g')\, dx && \text{We use FTC.} \\ f \cdot g &= \int (f' \cdot g + f \cdot g')\, dx \\ &= \int f' \cdot g\, dx + \int f \cdot g'\, dx . \end{aligned}$$

Now, these two integrals are very similar and either of them may be seen as the integral of a certain product. We derive something useful from this:

Theorem 2.9.1: Integration by Parts

For two integrable functions f and g, we have:

$$\boxed{\int f \cdot g'\, dx = f \cdot g - \int f' \cdot g\, dx}$$

We can also use the *substitution formula*,

$$dh = h'(x)\, dx .$$

To obtain a more compact version:

Corollary 2.9.2: Integration by Parts

For two integrable functions f and g, we have

$$\boxed{\int f\, dg = fg - \int g\, df}$$

The formula is traditionally restated with these, changed, names of the functions or variables:

$$\boxed{\int udv = uv - \int vdu}$$

When we are to decide which technique of integration to use, we recognize that for integration by part we have to see in the integrand:

- a multiplication, but
- no composition.

Then:

- The approach won't work for
$$\int xe^{x^2}\,dx\,,$$
but we have substitution for that.
- Nor for
$$\int e^{x^2}\,dx\,,$$
but we can look it up.
- The approach *might* work for
$$\int xe^{x}\,dx\,.$$

Example 2.9.3: split integrand

If we are to integrate this, we need to match it with the integral in the formula. These two must be equal:
$$\int u\,dv$$
$$\int x\cdot e^x\,dx.$$
Unfortunately, there are at least two ways to match them:
- (a) $u = e^x, \quad dv = x\,dx$, and
- (b) $u = x, \quad dv = e^x\,dx$.

We'll have to do both.

(a) To use the formula, we need the derivative of u and the integral of v:
$$\begin{aligned} u = e^x &\quad\Longrightarrow\quad du = e^x\,dx \\ dv = x\,dx &\quad\Longrightarrow\quad v = \int x\,dx = \frac{x^2}{2} \end{aligned}$$
Integrating to find v is the first *part* of integration. The second *part* is the one in the formula:
$$\int udv = uv - \int vdu = e^x\cdot\frac{x^2}{2} - \int \frac{x^2}{2}\cdot e^x\,dx\,.$$
Unfortunately, we discover that the new integral looks even complex than the original! Indeed, the power of x went up. Before attempting other techniques, let's try to reverse the choice of u and v.

(b) Once again, we need the derivative of u and the integral of v:
$$\begin{aligned} u = x &\quad\Longrightarrow\quad du = dx \\ dv = e^x\,dx &\quad\Longrightarrow\quad v = \int e^x\,dx = e^x \end{aligned}$$
We substitute these into the formula:
$$\int udv = uv - \int vdu = x\cdot e^x - \int e^x\,dx\,.$$
We pause here to stop and appreciate the fact that the new integral is so less complex than the original! That's because the power of x went down. We finish the computation:
$$\int xe^x\,dx = x\cdot e^x - \int e^x\,dx = xe^x - e^x + C\,.$$

The lesson seems to be:

- Choose for u the part of the integrand that will be simplified by differentiation.
- Choose for dv the part of the integrand that will be simplified by integration, or at least will remain as simple.

Example 2.9.4: split integrand

Integrate:

$$\int x^2 e^x \, dx \, .$$

Once again, there are (at least) two ways to choose u and dv:

- (a) $u = e^x, \quad dv = x^2 \, dx$, and
- (b) $u = x^2, \quad dv = e^x \, dx$.

We'll try both.

(a) We find the derivative of u and the integral of v:

$$\begin{array}{lll} u = e^x & \Longrightarrow & du = e^x \, dx \\ dv = x^2 \, dx & \Longrightarrow & v = \int x^2 \, dx = \dfrac{x^3}{3} \end{array}$$

Even though du is just as simple as u, integration of dv has made things worse. Indeed, we have:

$$\int u dv = uv - \int v du = e^x \cdot \frac{x^3}{3} - \int \frac{x^3}{3} \cdot e^x \, dx \, .$$

It's not simpler as the power of x goes up! We reverse the choice of u and v.

(b) The derivative of u and the integral of v:

$$\begin{array}{lll} u = x^2 & \Longrightarrow & du = 2x \, dx \\ dv = e^x \, dx & \Longrightarrow & v = \int e^x \, dx = e^x \end{array}$$

Here dv is simpler than u; that's a good sign. We substitute these into the formula:

$$\int u dv = uv - \int v du = x^2 \cdot e^x - \int e^x \cdot 2x \, dx \, .$$

Again, we pause to appreciate the fact that the integration task has been simplified! That's because the power of x went down. We finish the computation using the Integration by Parts formula and the result of the last example:

$$\int x^2 e^x \, dx = x^2 \cdot e^x - \int e^x 2x \, dx = x^2 e^x - 2xe^x + 2e^x + C \, .$$

The lesson is that integration by parts might bring simplification of the integral and might require another application of integration by parts.

Exercise 2.9.5

Apply the formula found in this example to the previous example, part (a).

Exercise 2.9.6

In the last example, try this decomposition of the integrand: $u = x, \; dv = xe^x$.

Example 2.9.7: recursion

Integrate:

$$\int x^3 \sin x \, dx\,.$$

There are two ways to split the integrand, u and dv, but we already know that integration by part will reduce the power x, if we choose $u = x^3$. We are left with $dv = \sin x$. Then

$$\begin{aligned} u = x^3 \quad &\Longrightarrow \quad du = 3x^2 \, dx \\ dv = \sin x \, dx \quad &\Longrightarrow \quad v = \int \sin x \, dx = -\cos x \end{aligned}$$

By parts:

$$\int x^3 \sin x \, dx = uv - \int v du = -x^3 \cos x - \int 3x^2 \cdot \sin x \, dx\,.$$

The last integral is almost identical to the original but the power of x is down by 1.

Exercise 2.9.8

Finish the computation by integrating by parts two more times.

Example 2.9.9: can't split

Integrate:

$$\int \cos^{-1} x \, dx\,.$$

There seems to be nothing to split in the integrand!.. There is:

$$\begin{aligned} u = \cos^{-1} x \quad &\Longrightarrow \quad du = -\frac{1}{\sqrt{1-x^2}} \, dx \\ dv = dx \quad &\Longrightarrow \quad v = \int dx = x \end{aligned}$$

By parts:

$$\begin{aligned} \int \cos^{-1} x \, dx &= uv - \int v du \\ &= \cos^{-1} x \cdot x - \int x \left(-\frac{1}{\sqrt{1-x^2}} \right) dx && \text{Done with parts.} \\ &= x \cos^{-1} x + \int \frac{x}{\sqrt{1-x^2}} \, dx && \text{By substitution:} \qquad z = 1 - x^2 \\ &= x \cos^{-1} x + \int \frac{1}{\sqrt{z}} \frac{dz}{-2} && \qquad \Rightarrow dz = -2x dx \\ &= x \cos^{-1} x - \frac{1}{2} z^{-1/2} \, dz \\ &= x \cos^{-1} x - \frac{1}{2} \frac{z^{1/2}}{1/2} + C && \text{Back-substitution.} \\ &= x \cos^{-1} x - \sqrt{1-x^2} + C. \end{aligned}$$

Exercise 2.9.10

Apply the Integration by Parts formula to the integral,

$$\int x e^x \, dx\,,$$

with these two choices of the "parts":
- (a) x and $e^x\,dx$,
- (b) e^x and $x\,dx$.

2.10. Approaches to integration

Let's summarize what we know about integration and compare it to what we know about differentiation.

First, the similarities.

Just as there is a list of elementary derivatives, we have a list of elementary integrals. In fact, the latter comes from the former.

Here they are:

$$\begin{array}{lll}
(x^s)' = sx^{s-1} & \displaystyle\int x^s\,dx = \frac{1}{s+1}x^{s+1}+C, & \text{for } s\neq -1\\
(\ln x)' = \dfrac{1}{x} & \displaystyle\int \frac{1}{x}\,dx = \ln x + C & \\
(e^x)' = e^x & \displaystyle\int e^x\,dx = e^x + C & \\
(\sin x)' = \cos x & \displaystyle\int \cos x\,dx = \sin x + C & \\
(\cos x)' = -\sin x & \displaystyle\int \sin x\,dx = -\cos x + C &
\end{array}$$

It's just a short list of integral formulas for specific functions. So, given a function, we find it on the list and, automatically, its integral, just like with the derivatives. In either case, the list is very short!

There are differences already between the two. The presence of "$+C$" in each integral indicates that the answer contains infinitely many functions. Also, the formulas for integrals only remain valid when limited to *intervals*.

Just as there are algebraic rules of differentiation, there are algebraic rules of integration. The latter comes, in part, from the former.

$$\begin{array}{rll}
\text{SR} & (f+g)' = f'+g' & \displaystyle\int (f+g)\,dx = \int f\,dx + \int g\,dx\\
\text{CMR} & (cf)' = cf' & \displaystyle\int (cf)\,dx = c\int f\,dx\\
\text{LCR} & (f(mx+b))' = mf'(mx+b) & \displaystyle\int f(mx+b)\,dx = \frac{1}{m}\int f(t)\,dt\Big|_{t=mx+b}
\end{array}$$

The way we apply these rules is very similar to the one for derivatives:

- When the function to be integrated is the sum of two (just as when it was to be differentiated), we split the integral into two and then deal with either (simpler) integral separately.
- When the function to be integrated is a constant multiple of another (just as when it was to be differentiated), we factor it out and then deal with the remaining (simpler) integral.

The similarities stop here!

What about the *Product Rule* for integration? There is none in the above sense.

> ▶ When the function to be integrated is the product of two (unlike when it was to be differentiated), we *can't* split the integral into two and then deal with either integral separately.

The reason is that the Product Rule for differentiation can't be easily reversed... unless one of the functions is, in fact, the derivative of a function that we know or can find. That's the *Integration by Parts* formula:

$$\boxed{\int fg'\,dx = fg - \int gf'\,dx}$$

And what about the *Quotient Rule* for integration? There is none, unless you are willing to interpret division as multiplication by the reciprocal.

Now, what about the *Chain Rule*? Same problem as with the products:

> ▶ When the function to be integrated is the composition of two (unlike when it was to be differentiated), we *can't* split the integral into two and then deal with either integral separately.

The reason is that the Chain Rule for differentiation can't be easily reversed... unless the derivative of the function on the inside is, in fact, present as a factor. That's the *Integration by Substitution* formula:

$$\boxed{\int (f\circ g)\cdot g'\,dx = \int f\,du}$$

Example 2.10.1: routine

Compute:

$$\int_0^1 \left(x^3 + 3e^x - \sin x\right)\,dx\,.$$

Ignore the bounds at first:

$$\begin{aligned}
\int \left(x^3 + 3e^x - \sin x\right)\,dx &\ \overset{\text{SR}}{=\!=\!=}\ \int x^3\,dx + \int 3e^x\,dx + \int \sin x\,dx\\
&\ \overset{\text{CMR}}{=\!=\!=}\ \int x^3\,dx + 3\int e^x\,dx + \int \sin x\,dx\\
&\ \overset{\text{formulas}}{=\!=\!=}\ \frac{x^4}{4} + 3\cdot e^x - (-\cos x) + C\\
&\ \overset{\text{simplify}}{=\!=\!=}\ \frac{1}{4}x^4 + 3e^x + \cos x + C\,.
\end{aligned}$$

That's the hard part, finding antiderivatives. Now the easy part:

$$\begin{aligned}
\int_0^1 \left(x^3 + 3e^x - \sin x\right)\,dx &\ \overset{\text{FTC}}{=\!=\!=}\ \frac{1}{4}x^4 + 3e^x + \cos x\Big|_0^1\\
&\ \overset{\text{substitute}}{=\!=\!=}\ \left(\frac{1}{4}1^4 + 3e^1 + \cos 1\right) + \left(\frac{1}{4}0^4 + 3e^0 + \cos 0\right)\\
&\ =\!=\!=\ \frac{1}{4} + 3e + \cos 1 - 0 - 3 - 1\,.
\end{aligned}$$

The hard part is easy when there is no multiplication, division, or composition of functions.

Warning!

The most important approach to integration is using the table of integrals!

The main difference is that *differentiation never fails* but integration may fail in the sense that the integral might turn out to be a function we have never seen before or even a function that no-one has seen before!

Example 2.10.2: define functions as integrals

We can use this idea to re-discover "familiar" functions – starting at the other end.

For example, integrating this rational function produces the logarithm:

$$\frac{1}{t} \quad \text{leading to} \quad \int_1^x \frac{1}{t}\, dt = \ln x\,.$$

We have thus defined the logarithm without reference to the inverse of the exponential function.

Also, integrating this algebraic function produces the arcsine:

$$\frac{1}{\sqrt{1-t^2}} \quad \text{leading to} \quad \int_0^x \frac{1}{\sqrt{1-t^2}}\, dt = \sin^{-1} x\,.$$

We have defined the arcsine without reference to the inverse of the sine.

2.11. *The areas of infinite regions: improper integrals*

We can only compute definite integrals over *bounded* intervals such as $[a,b]$.

Example 2.11.1: what happens to antiderivative

Consider again this deceptive formula:

$$\int \frac{1}{x}\, dx \overset{???}{=\!=\!=} \ln|x| + C,\ x \neq 0\,.$$

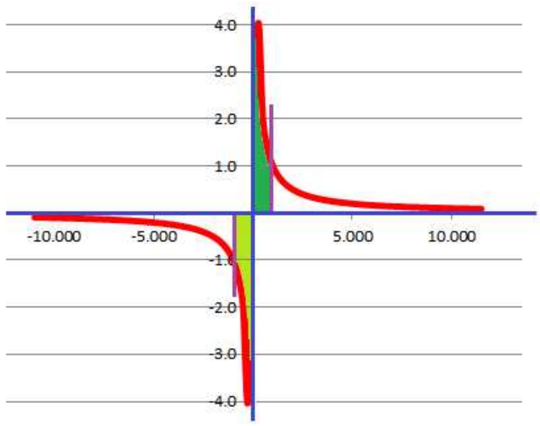
The formula is false as stated! It holds only on either of *the two intervals*, separately, of the domain of $\frac{1}{x}$, i.e., $(-\infty,0)$ and $(0,\infty)$, but not on the set $(-\infty,0)\cup(0,\infty)$. This means that C can vary from the one to the other! In fact, this is the antiderivative:

$$F(x) = \begin{cases} \ln|x| + C & \text{for } x < 0\,, \\ \ln|x| + K & \text{for } x > 0\,. \end{cases}$$

It is even more dangerous to ignore the gap in the domain when we deal with definite integrals. For example, one might produce this from the formula:

$$\int_{-1}^{1} \frac{1}{x}\, dx \overset{???}{=\!=\!=} \ln|x| \Big|_{-1}^{1} = \ln 1 - \ln|-1| = 0\,.$$

This is untrue because the integral is indeterminate as a limit. Indeed, f is not integrable on $[-1,1]$ simply because it's undefined at $x=0$. Furthermore, even though the positive and the negative areas seem to cancel each other, this is false because both are, in fact, *infinite*:



We shouldn't becasual about doing algebra with infinities (Volume 2):

$$\infty - \infty \overset{???}{=\!=\!=} 0\,.$$

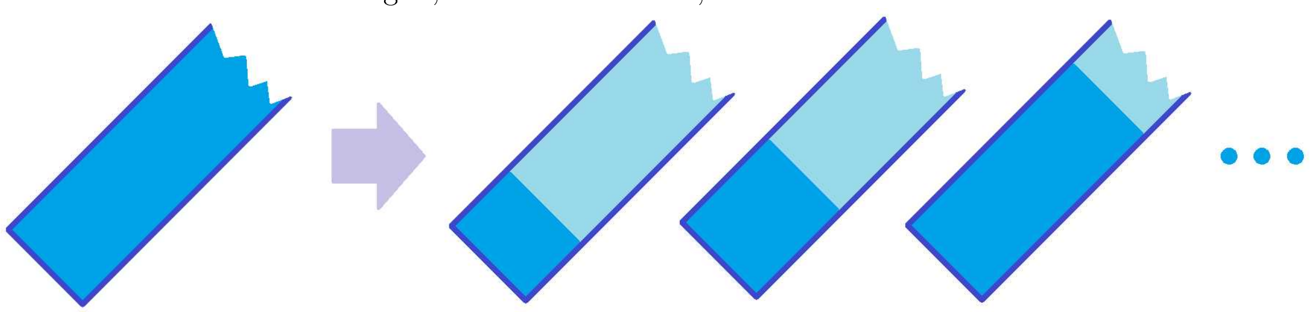
Exercise 2.11.2

Show that the function in the example won't become integrable whatever number we assign to $x = 0$.

We will next try to understand the meaning of the area of an infinite or, better, *unbounded* region.

Example 2.11.3: infinite bottle

The area of an "infinite rectangle", like the one below, must be infinite:



Why or in what sense? This region contains a growing sequence of (real, finite) rectangles the areas of which grow to infinity. In other words, it would take an infinite amount of water to fill such a bottle.

We have come to understand the meaning of the areas of (some) bounded regions. This will be our approach:

- "Exhaust" an unbounded region with a sequence of bounded regions.
- Find their areas.
- Examine the *limit* of these areas.

It's no different, in principle, from exhausting a circle with polygons.

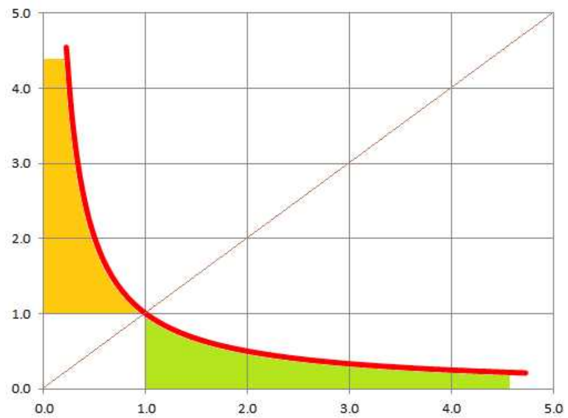
We will restrict our attention to regions:

1. unbounded with respect to the x-axis (infinitely wide), and
2. unbounded with respect to the y-axis (infinitely tall).

As we only deal with regions determined by graphs of functions,

1. the former case is about functions with *unbounded domains* or, better, unbounded domains of integration, and
2. the latter about functions with *unbounded ranges* (i.e., unbounded functions).

Even though these two classes of functions are very different, the issue is the same. For example, $y = 1/x$ defines two *identical* unbounded regions:



We start with the former case: *an unbounded domain of integration.*

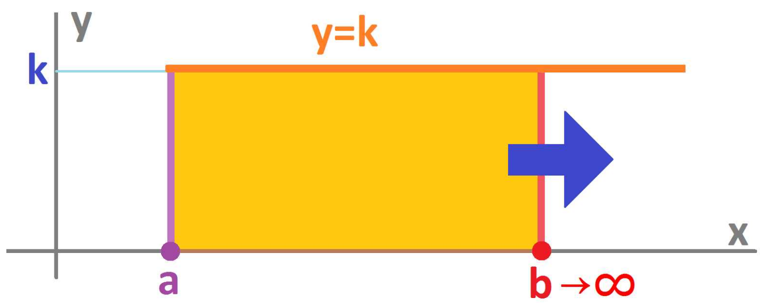
Example 2.11.4: constant

Consider a constant function,

$$f(x) = k \text{ on } [a, \infty),\ k > 0\,.$$

Then, the area of the region above the interval $[a, b]$ with $b > a$ (and the integral from a to b) is equal to $(b - a)k$. Furthermore,

$$\text{Area } = (b - a)k \ \to +\infty \ \text{ as } \ b \to +\infty\,.$$



Therefore, the area of the infinite strip is infinite, as expected.

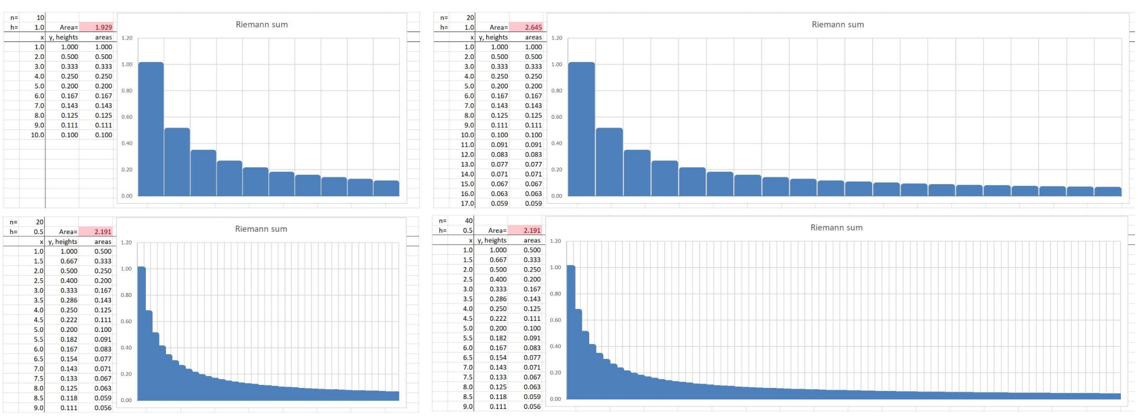
Example 2.11.5: reciprocal

Consider again

$$f(x) = 1/x \ \text{ on } [1, \infty)\,.$$

The area under this graph over this ray in the x-axis is shown above.

This unbounded region is exhausted by bounded ones. How? The obvious approach is to use the Riemann sums. The challenge is that one has to both make the rectangles thinner and thinner (as before) and make the right end of the interval extend more and more to the right.



(Note that when $h = 1$, this sum is called a *series* to be discussed in Chapter 5.)

An alternative approach is to rely on what we already know about areas under the graphs when the interval is bounded. The underlying ray of this region is exhausted with bounded intervals. They all have the same left bound, 1, but the right bound, b, is approaching infinity:

$$[1, b] \quad \text{leading to} \quad [1, +\infty) \ \text{ as } b \to +\infty\,.$$

Then,

$$\begin{aligned} \int_1^b \frac{1}{x}\,dx &= \ln b - \ln 1 \\ &\to \infty \qquad\qquad \text{as } b \to +\infty\,. \end{aligned}$$

The area of the unbounded region is infinite.

Initially, we will concentrate on *rays* as the domains:

$$(-\infty, b] \ \text{ and } \ [a, +\infty)\,.$$

The rays will have to be "exhausted" with *closed and bounded* intervals, on each of which we face the usual Riemann integrals.

Example 2.11.6: more reciprocal powers

We discovered that the band under the graph of $y = 1/x$ is narrowing down but not fast enough to avoid growing its area to infinity. A function with a narrower strip won't necessarily avoid having an infinite area: $y = 1/(3x)$. Instead, let's try, in contrast, a function that is *decreasing much faster*. How about $y = 1/x^2$? We have:

$$\begin{aligned} \int_1^b \frac{1}{x^2}\,dx &= -\frac{1}{b} + 1 \\ &\to 1 \qquad\qquad \text{as } b \to +\infty\,. \end{aligned}$$

So, 1 is the area of the unbounded region. It's finite!

In addition to the areas, integrals can also be explained in terms of *motion*. If our velocity is positive but declining as $1/x^2$, where x is time, the distance we will cover over an infinite period of time will not be infinite!

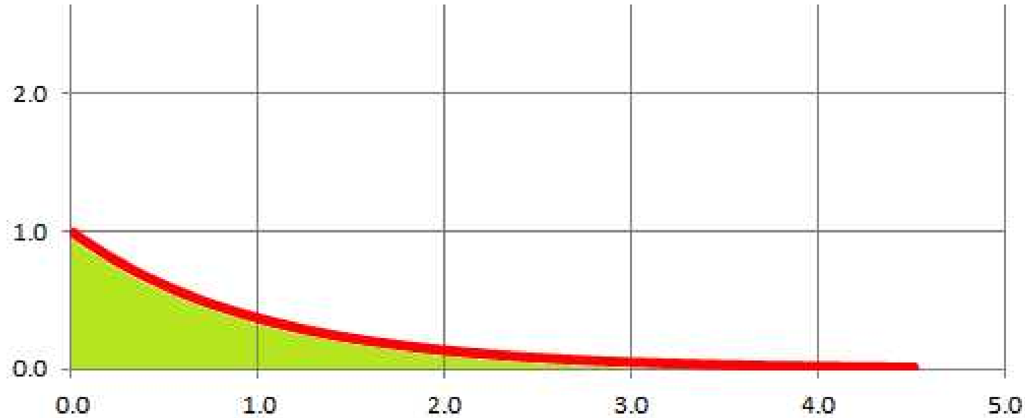
Example 2.11.7: exponential

What function decreases faster than all $1/x^n$, $n = 1, 2, 3, ...$? It's the *exponential decay function*:

$$f(x) = e^{-x} \text{ on } [1, \infty)\,.$$

Again, the unbounded region under this graph above this ray is exhausted by exhausting the underlying ray with the bounded intervals:

$$[1, b] \quad \text{leading to} \quad [1, +\infty) \text{ as } b \to +\infty\,.$$



Then,

$$\begin{aligned} \text{total area} &= \lim_{b\to+\infty} \int_1^b e^{-x}\, dx \\ &= \lim_{b\to+\infty} (-e^{-b} - (-e^{-1})) \\ &= \frac{1}{e}\,. \end{aligned}$$

The main discovery is that *the area of an unbounded region doesn't have to be be infinite*!

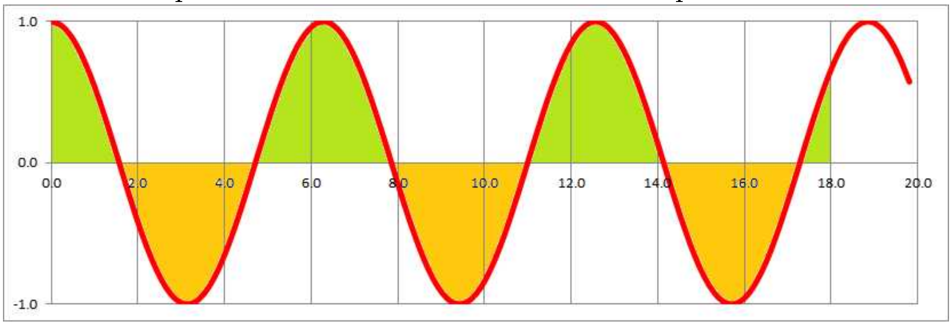
Exercise 2.11.8

Find the area under the graphs of the following over $[1,\infty)$:

1. $y = \frac{1}{x^2}$
2. $y = \frac{1}{\sqrt{x}}$

Example 2.11.9: sinusoid

What if the function isn't all positive? What is the area – over the positive numbers – under a sinusoid?



The analysis is identical:

$$\text{total area } = \lim_{b\to+\infty} \int_0^b \cos x\, dx = \lim_{b\to+\infty} (\sin b - \sin 0) \quad \text{DNE}\ .$$

The limit doesn't exist and, therefore, there is no area.

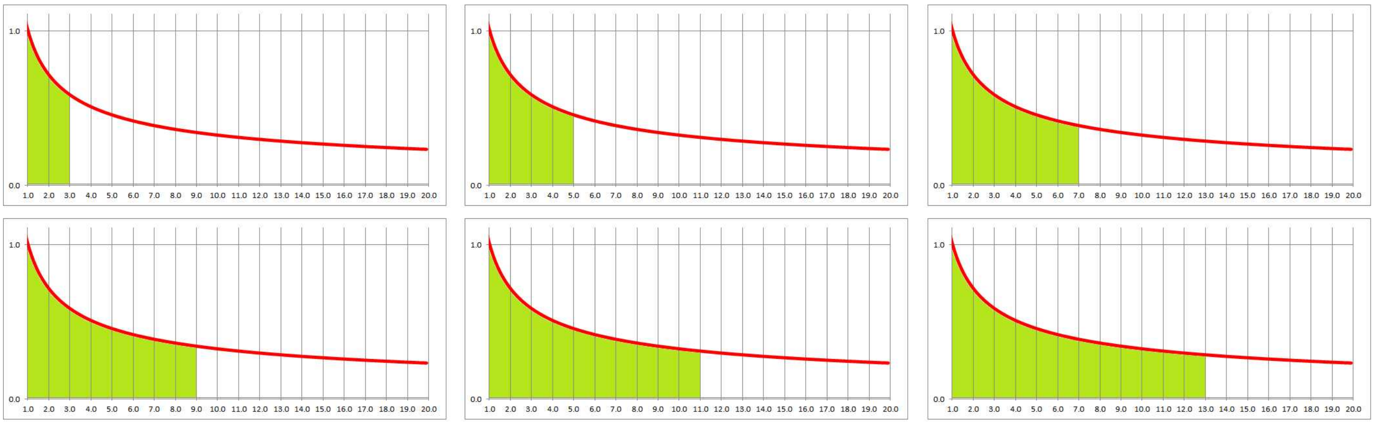
Warning!

"No area" isn't the same as "zero area".

Just as with all limits, there are three possible outcomes: This may be a number, or it may be infinite, or it may be undefined.

This is the summary.

We "exhaust" the unbounded domain of integration, $(-\infty, b]$ or $[a,\infty)$, with bounded ones. If the function is integrable on these closed and bounded domains, we then "exhaust" a possibly infinite area of over this domain with finite ones:



If the limit of this integral exists, it is denoted as follows:

Integral over ray

$$\int_{-\infty}^{b} f(t)\,dt = \lim_{a\to-\infty}\int_{a}^{b} f(t)\,dt$$

and

$$\int_{a}^{\infty} f(t)\,dt = \lim_{b\to+\infty}\int_{a}^{b} f(t)\,dt$$

These limits, of course, can be infinite.

Warning!

Even though the notation suggests that the domain of integration is the whole ray, this is nothing but an abbreviation for the limit on the right.

We also define the integral over the whole real line $(-\infty,\infty)$ in terms of the ones over rays, as the sum of two corresponding integrals (limits) with the following notation:

Integral over line

$$\int_{-\infty}^{\infty} f(t)\,dt = \int_{-\infty}^{0} f(t)\,dt + \int_{0}^{\infty} f(t)\,dt$$

In the case of infinite limits, we utilize the *algebra of infinities* as shortcuts (Volume 2):

$$\begin{array}{llll}
(\text{ number }) & + & (+\infty) & = +\infty \\
(\text{ number }) & + & (-\infty) & = -\infty \\
(+\infty) & + & (+\infty) & = +\infty \\
(-\infty) & + & (-\infty) & = -\infty
\end{array}$$

Example 2.11.10: exponential

Let's compute the integral,

$$\begin{aligned}\int_{-\infty}^{+\infty} e^x\,dx &= \int_{-\infty}^{0} e^x\,dx + \int_{0}^{\infty} e^x\,dx\\ &= \lim_{a\to-\infty}\int_{a}^{0} e^x\,dx + \lim_{b\to+\infty}\int_{0}^{b} e^x\,dx\\ &= \lim_{a\to-\infty}(1-e^a) + \lim_{b\to+\infty}(e^b-1)\,dx\,.\end{aligned}$$

The first limit is 1 but the second limit is infinite; therefore, our integral is infinite.

Exercise 2.11.11

Show that replacing 0 in the definition of the integral over $(-\infty, +\infty)$ with any real c will produce the same result.

Exercise 2.11.12

Show that replacing the last definition with

$$\int_{-\infty}^{\infty} f(t)\,dt \stackrel{?}{=} \lim_{R\to\infty}\int_{-R}^{R} f(t)\,dt\,,$$

won't always produce the same result.

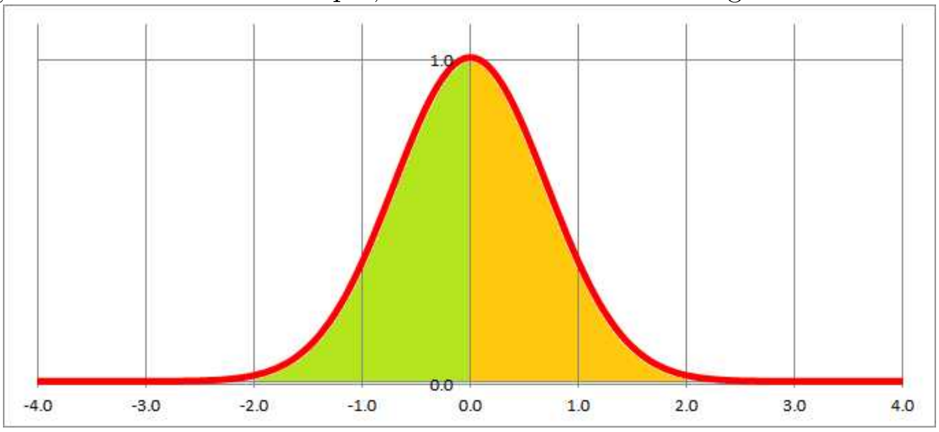
Definition 2.11.13: improper integrals on infinite intervals

The limits of the integrals above are called *improper integrals*. When such a limit exists, or the two limits in the last case exist, we say that the improper integral *converges* and the function is *integrable*; otherwise it *diverges*.

The latter terminology is borrowed from limits and it will be used again in Chapter 5.

Example 2.11.14: bell curve

The definition of the integral over $(-\infty, \infty)$ follows the additivity of the integral that comes from the idea of additivity of the areas. For example, this is what such an integral looks like:



The function is e^{-x^2} and the integral is known to be convergent.

Some integrals should be computed ahead of time.

Theorem 2.11.15: Improper Integrals of Reciprocals I

For any $a > 0$, we have

$$\int_a^\infty \frac{1}{x^p}\,dx = \begin{cases} \dfrac{a^{1-p}}{p-1} & \text{when } p > 1, \\ \infty & \text{when } 0 < p \leq 1. \end{cases}$$

Proof.

For $p \neq 1$, we have:

$$\begin{aligned} \int_a^\infty \frac{1}{x^p}\,dx &= \lim_{b\to\infty} \int_a^b \frac{1}{x^p}\,dx && \text{According to definition.} \\ &= \lim_{b\to\infty} \int_a^b x^{-p}\,dx && \text{Use PF next.} \\ &= \lim_{b\to\infty} \frac{1}{-p+1} x^{-p+1} \Bigg|_a^b \\ &= \frac{1}{-p+1} \lim_{b\to\infty} \left(b^{-p+1} - a^{-p+1}\right) \\ &= \frac{1}{-p+1} \left(\lim_{b\to\infty} b^{-p+1} - a^{-p+1}\right). \end{aligned}$$

The remaining limit is 0 when $-p+1 < 0$, and it is infinite when $-p+1 > 0$.

Exercise 2.11.16

Finish the proof.

In other words, the integral

- converges when $p > 1$,
- diverges when $0 < p \leq 1$.

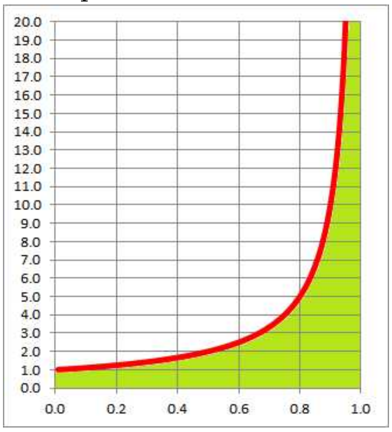
Now, the latter case: *unbounded functions and bounded domains of integration.*

Example 2.11.17: infinite area

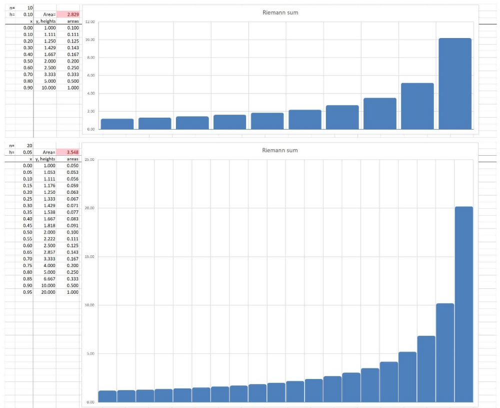
Consider

$$f(x) = \frac{1}{1-x} \quad \text{on } [0,1)\,.$$

The area under this graph over this half-open interval in the x-axis is shown below:



How do we understand the area under this graph over this interval? We can use the Riemann sums, exactly as always, as long as 1 *is not among its sample points*:



Alternatively, we take a hint from the analysis of case 1: This unbounded region is exhausted by bounded ones. How? The underlying interval is exhausted with closed intervals. They all have the same left bound, a, but the right bound, b, is approaching 1:

$$[0, b] \quad \text{leading to} \quad [0, 1) \text{ as } b \to 1\,.$$

Then,

$$\begin{aligned} \int_0^b \frac{1}{1-x}\, dx &= -\ln(1-b) - \ln 1 \\ &\to \infty \qquad\qquad\qquad \text{as } b \to 1\,, \end{aligned}$$

which is then the area of the unbounded region.

The definition of the Riemann integral simply doesn't apply to a function undefined at one of the ends of the interval. Instead, we consider a *restriction* of the function to a smaller, but closed, interval.

Initially, we will concentrate on *half-open intervals*:

$$(c, b] \quad \text{and} \quad [a, c)\,.$$

As you see, the analysis is very similar to the former case, with this substitution:

$$[a, \infty) \longrightarrow [a, c)\,.$$

Just as the former, the latter will have to be "exhausted" with *closed and bounded* intervals.

Example 2.11.18: finite area

Let's consider

$$f(x) = \frac{1}{\sqrt{1-x}} \quad \text{on } [0, 1)\,.$$

Even though their graphs look almost identical, this one increases slower than the last one. Again, the unbounded region under this graph over this half-open interval is exhausted by exhausting the underlying ray with the closed intervals: $[0, b]$ as $b \to 1$. Then we have:

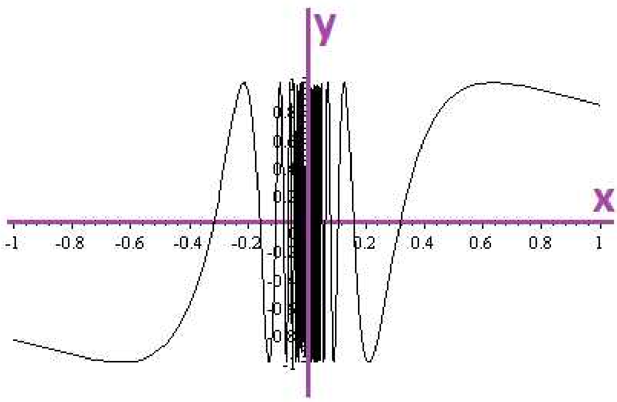
$$\begin{aligned} \text{Area of unbounded region} &= \lim_{b\to 1} \int_0^b \frac{1}{\sqrt{1-x}}\, dx \\ &= \lim_{b\to 1} -2\sqrt{1-x}\,\Big|_0^b \\ &= 2\,. \end{aligned}$$

The area of an unbounded region may be finite!

Example 2.11.19: divergence

What if the function isn't all positive? What is the area under an oscillating graph, such as this?

$$y = \sin \frac{1}{x}$$

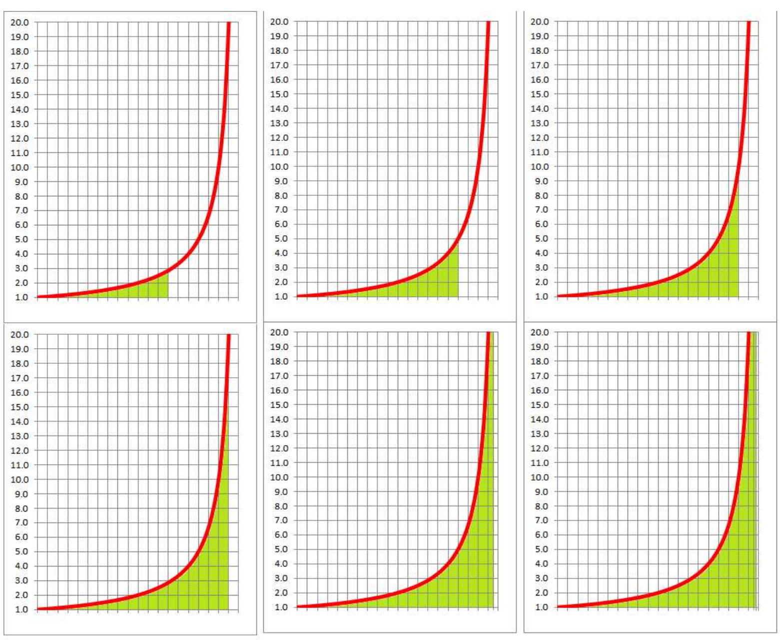


With the graph like this, one can *guess* that the limit doesn't exist and, therefore, there is no area.

Just as with all limits, there are three possible outcomes for these areas: a number, infinity, or undefined.

This is the summary.

We "exhaust" the half-open domain of integration, $(a, b]$ or $[a, b)$, with closed ones. If the function is integrable on these closed and bounded domains, we then "exhaust" a possibly infinite area over this domain with finite ones.



If the limit of this integral exists, it is denoted as follows:

Integral over half-open intervals

$$\int_c^b f(t)\,dt = \lim_{a\to c} \int_a^b f(t)\,dt$$

and

$$\int_a^c f(t)\,dt = \lim_{b\to c} \int_a^b f(t)\,dt$$

These limits, of course, can be infinite.

Warning!

The notation is unfortunately identical to the one for proper integrals, but this is nothing but an abbreviation for the limit on the right.

We also define the integral over an open interval in terms of the ones over half-open intervals, with the following notation for any c between a and b:

Integral over open interval

$$\int_a^b f(t)\,dt = \int_a^c f(t)\,dt + \int_c^b f(t)\,dt$$

In the case of infinite limits, we follow the algebra of infinities above.

For case 2, we *repeat* the definition for case 1.

Definition 2.11.20: improper integrals on finite intervals

The limits of the integrals above are (also) called *improper integrals*. When the limit exists, or the two limits in the last case exist, we say that the improper integral *converges* and the function is *integrable*; otherwise it *diverges*.

Exercise 2.11.21

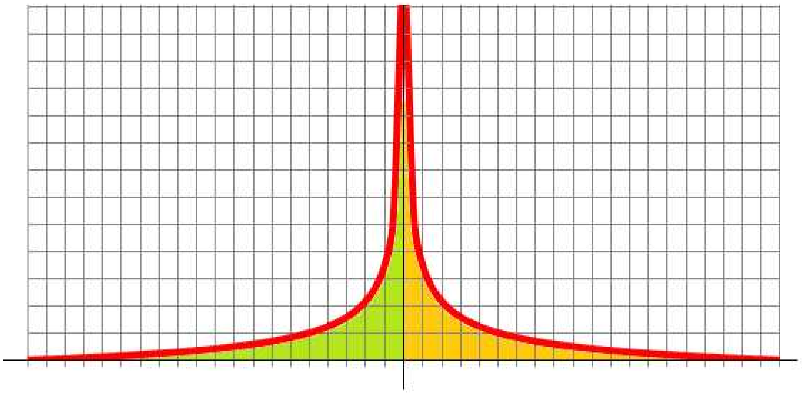
Show that replacing the last definition with

$$\int_a^b f(t)\,dt \overset{?}{=} \lim_{\varepsilon\to 0^+} \int_{a+\varepsilon}^{b-\varepsilon} f(t)\,dt\,.$$

won't produce the same result.

Example 2.11.22: hole in domain

The definition of the integral over an interval with a possible missing point inside follows additivity of the integral that comes from the idea of additivity of the areas. For example, this is what such an integral looks like:



The function is

$$\frac{1}{\sqrt{|x|}}\,,$$

and the integral is known to be convergent.

Theorem 2.11.23: Improper Integrals of Reciprocals II

For any $a > 0$, we have

$$\int_0^b \frac{1}{x^p}\,dx = \begin{cases} \dfrac{b^{1-p}}{1-p} & \text{when } 0 < p < 1, \\ \infty & \text{when } p \geq 1. \end{cases}$$

Proof.

For $p \neq 1$, we have:

$$\begin{aligned} \int_0^b \frac{1}{x^p}\,dx &= \lim_{a\to 0^+} \int_a^b \frac{1}{x^p}\,dx && \text{By definition.} \\ &= \lim_{a\to 0^+} \int_a^b x^{-p}\,dx && \text{We use PF next.} \\ &= \lim_{a\to 0^+} \frac{1}{-p+1} x^{-p+1} \Bigg|_a^b \\ &= \frac{1}{-p+1} \lim_{a\to 0^+} \left(b^{-p+1} - a^{-p+1}\right) \\ &= \frac{1}{-p+1} \left(b^{-p+1} - \lim_{a\to 0^+} a^{-p+1}\right). \end{aligned}$$

The remaining limit is 0 when $-p+1 > 0$, and it is infinite when $-p+1 < 0$.

Exercise 2.11.24

Finish the proof.

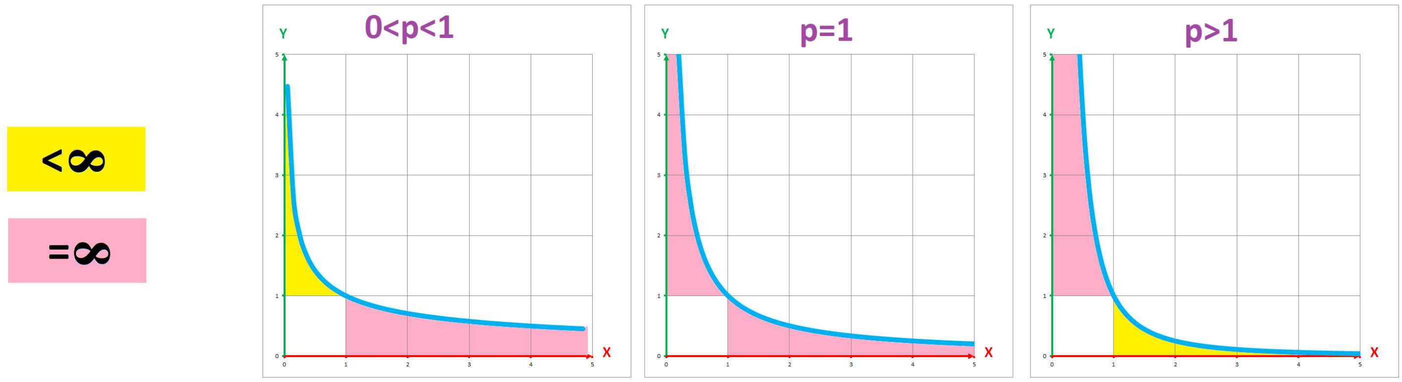
Exercise 2.11.25

Match the integrals and the areas of the two theorems about integrals of the reciprocals. Hint: It's about symmetry.

In other words, the integral $\displaystyle\int_0^b \frac{1}{x^p}\,dx$:

- converges when $p > 1$,
- diverges when $0 < p \leq 1$.

This is the summary of the two types of improper integrals for these functions:



Exercise 2.11.26

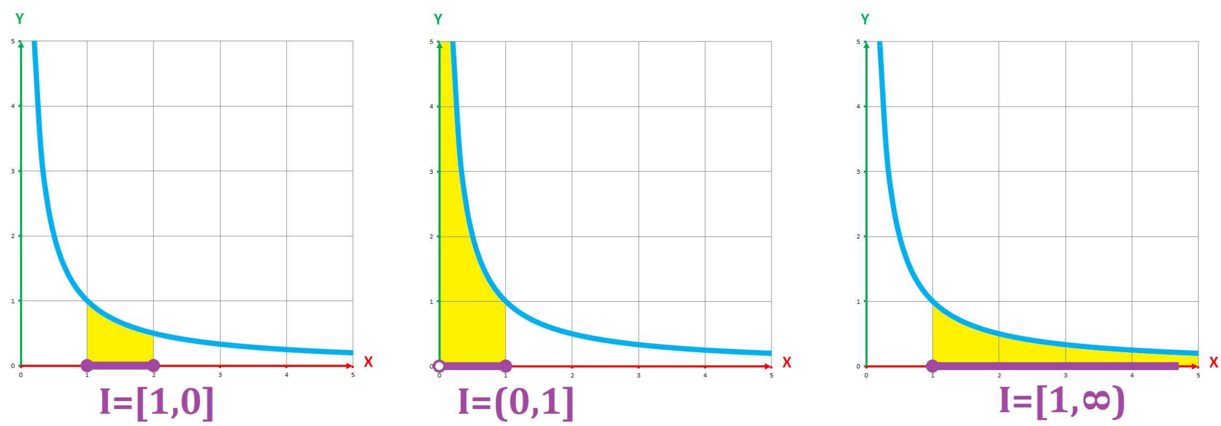
What possible values can the area between the graph of a function and its asymptote take? Give an example for each value.

2.12. Properties of proper and improper definite integrals

Thus, we have extended the idea of Riemann integral with the domain of integration:

- closed bounded intervals, such as $[a, b]$, to
- half-open, such as $(a, b]$ and $[a, b)$, and also possibly infinite, such as $(-\infty, b]$ and $[a, \infty)$, and further to
- open intervals, such as (a, b), and possibly infinite, such as $(-\infty, +\infty)$.

It is possible to have a uniform treatment of these cases:



We outline it below.

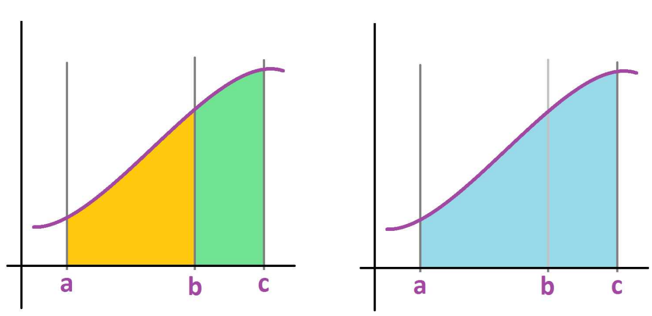
If we denote an interval by I, all these integrals can be written in the same notation:

Integral over interval

$$\int_I f\, dx$$

These integrals have identical properties. In fact, the properties of improper integrals follow from the corresponding properties of proper integrals, which in turn come from the properties of limits.

As regions are joined together via union, their areas are added – even though the regions may be unbounded. The area interpretation of additivity is the same as before, as long as the integrals are convergent:



Theorem 2.12.1: Additivity of Integrals

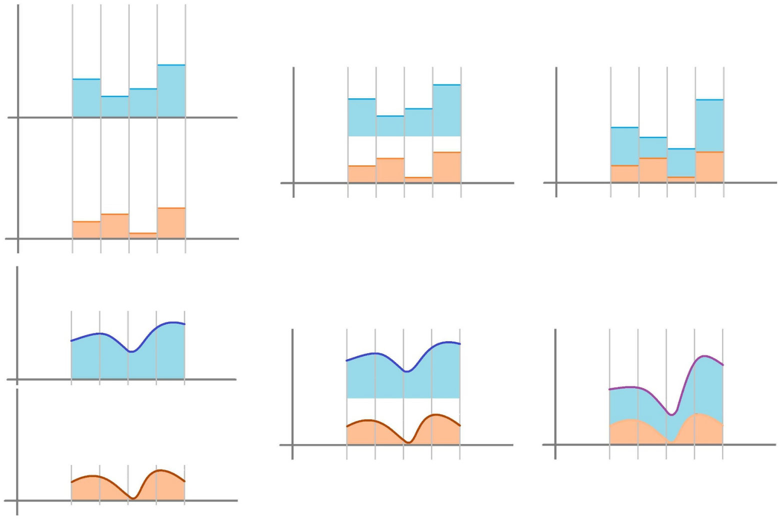
Suppose f is integrable over intervals I and J that share one point. Then f is integrable over $I \cup J$, and we have:

$$\boxed{\int_I f\,dx + \int_J f\,dx = \int_{I\cup J} f\,dx}$$

Theorem 2.12.2: Integrability

If f is integrable over interval I, then it is also integrable over any interval $J \subset I$.

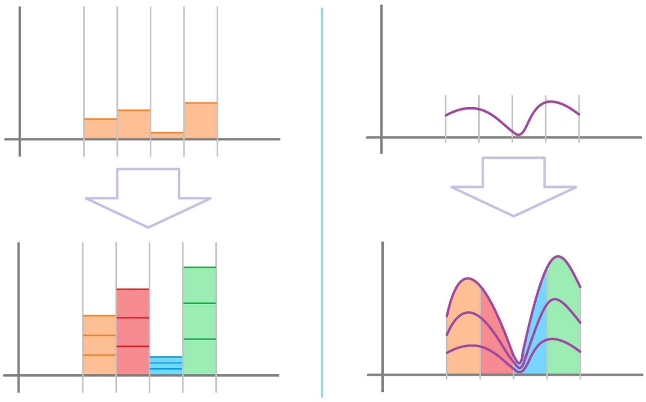
The algebraic rules are also the same.



Theorem 2.12.3: Sum Rule for Integrals

Suppose f and g are integrable functions over interval I. Then so is $f+g$, and we have:

$$\boxed{\int_I (f+g)\,dx = \int_I f\,dx + \int_I g\,dx}$$



Theorem 2.12.4: Constant Multiple Rule for Integrals

Suppose f is an integrable function over interval I. Then so is $c \cdot f$ for any real

c, and we have:

$$\int_I (c \cdot f)\, dx = c \cdot \int_I f\, dx$$

Exercise 2.12.5

Prove the two theorems. Hint: Use the rules of limits.

Exercise 2.12.6

What is the Fundamental Theorem of Calculus for improper integrals?

The main part of the construction of an improper integral is one of these limits:

$$\lim_{x\to b} \int_a^x \, dx \quad \text{and} \quad \lim_{x\to\infty} \int_a^x \, dx\,.$$

The convergence of the integral is the convergence of the limit. There is a way to predict what happens without evaluating the limit.

The following is a major theorem from Volume 2 (Chapter 2DC-1):

Theorem 2.12.7: Monotone Convergence

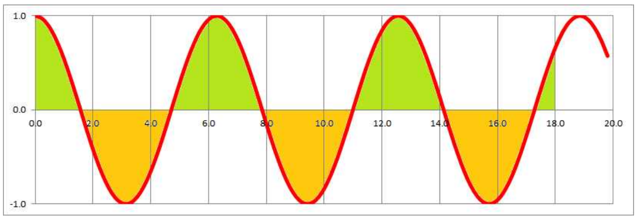
Every monotone and bounded sequence converges. Every monotone and bounded function converges at infinity.

We also know from Volume 2:

Theorem 2.12.8: Monotonicity Theorem

An antiderivative of a function is monotone when the function is either all positive or all negative.

Let's exclude the rest of the functions, such as sin and cos:



Then, we have the following:

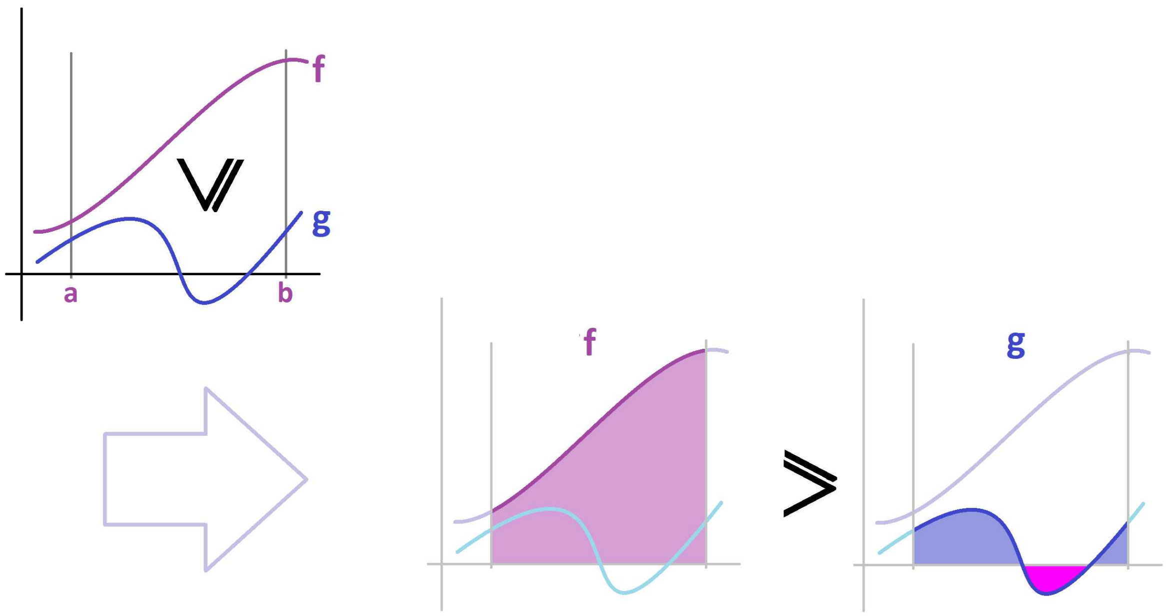
Theorem 2.12.9: Convergence of Non-negative Integral

If a function is non-negative, its integrals are either convergent or infinite.

Exercise 2.12.10

Prove the theorem.

Then, to establish convergence, we can use a direct comparison with another function, a function that has a convergent integral. The other function should be larger:



Similarly, to establish divergence, we can use a direct comparison with another function, a function that has a divergent integral. The other function should be smaller.

Example 2.12.11: comparison

Consider these two facts:

$$\text{The integral } \int_1^\infty \frac{1}{x^{1/2}}\,dx \quad \text{diverges} \quad \text{because} \quad \int_1^\infty \frac{1}{x^{1/3}}\,dx \quad \text{diverges.}$$

$$\text{The integral } \int_1^\infty \frac{1}{x^{3}}\,dx \quad \text{converges} \quad \text{because} \quad \int_1^\infty \frac{1}{x^{2}}\,dx \quad \text{converges.}$$

These conclusions come from the inequalities below:

$$\begin{array}{ccccccccc}
1/3 & \le 1/2 & \le & 2 & \le 3 & \Longrightarrow \\
x^{1/3} & \le x^{1/2} & \le & x^2 & \le x^3 & \Longrightarrow \\
\dfrac{1}{x^{1/3}} & \ge \dfrac{1}{x^{1/2}} & \ge & \dfrac{1}{x^2} & \ge \dfrac{1}{x^3} & \Longrightarrow
\end{array}$$

$$\int_1^\infty \frac{1}{x^{1/3}}\,dx \ \ge \int_1^\infty \frac{1}{x^{1/2}}\,dx \ = \infty > \int_1^\infty \frac{1}{x^2}\,dx \ \ge \int_1^\infty \frac{1}{x^3}\,dx$$

Exercise 2.12.12

What does the middle inequality give us?

The idea applies to all improper integrals:

Theorem 2.12.13: Comparison for Improper Integrals

Suppose I is an interval, and

$$0 \le f(x) \le g(x)$$

for all x in I. Then, for improper integrals over I, we have:

- *If the improper integral of f diverges, then so does the improper integral of g.*
- *If the improper integral of g converges, then so does the improper integral*

of f and we have:

$$0 \le \int_I f\, dx \le \int_I g\, dx\,.$$

Exercise 2.12.14

Prove the theorem.

Suppose the functions are non-negative. According to the *Convergence Theorem*, the following notation makes sense:

Convergence of improper integral

$$\int_I f\, dx < \infty$$

Divergence of improper integral

$$\int_I f\, dx = \infty$$

Then the *Comparison Theorem* above can be read from these simple inequalities:

$$\int_I f\, dx \ge \int_I g\, dx = \infty$$

and

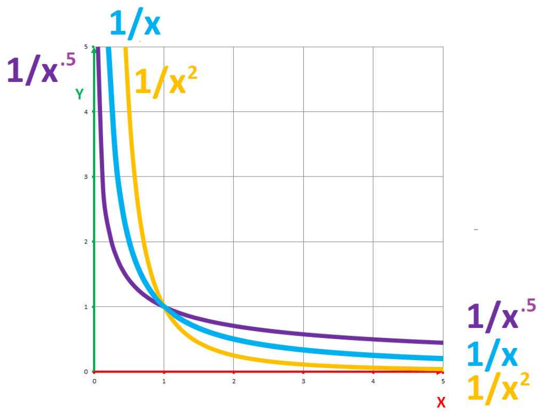
$$\int_I f\, dx \le \int_I g\, dx < \infty\,.$$

Exercise 2.12.15

What if we use *strict* inequalities in the above example?

Exercise 2.12.16

What can we derive about the convergence/divergence of the improper integrals of the reciprocal powers based entirely on that of $1/x$? Hint:



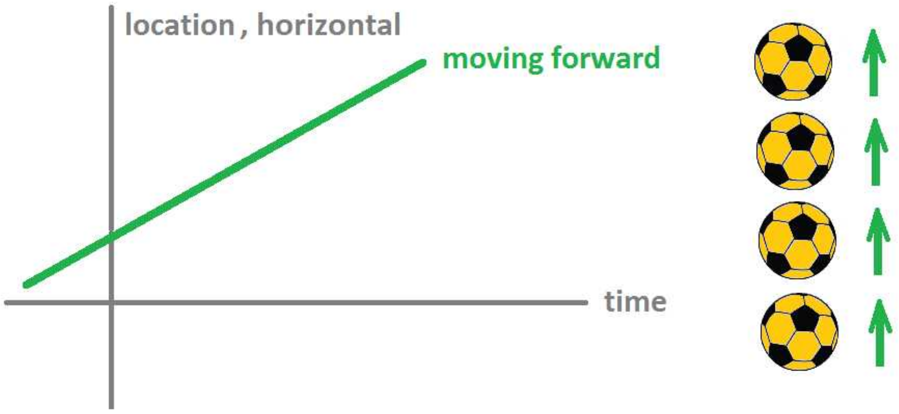
Exercise 2.12.17

State and prove the Squeeze Theorem for improper integrals.

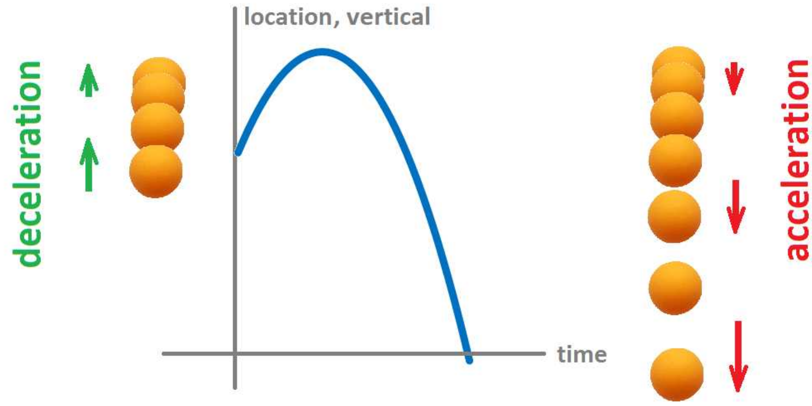
2.13. Shooting a cannon

Let's review what we know about free flight from Volume 2.

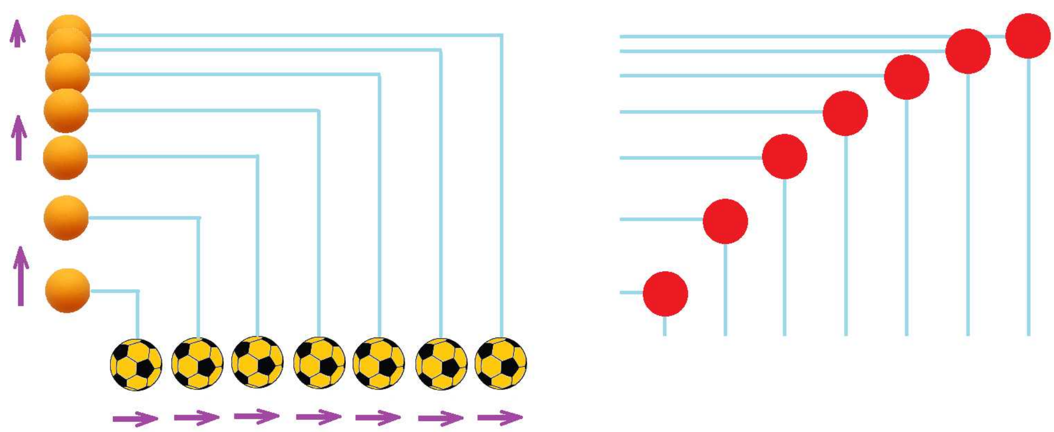
A soccer ball rolling on a horizontal plane will have a constant velocity:



A ping-pong ball thrown up in the air goes up, slows down until it stops for an instant, and then accelerates toward the surface:

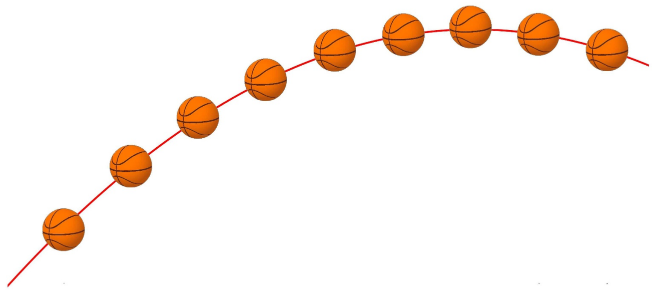


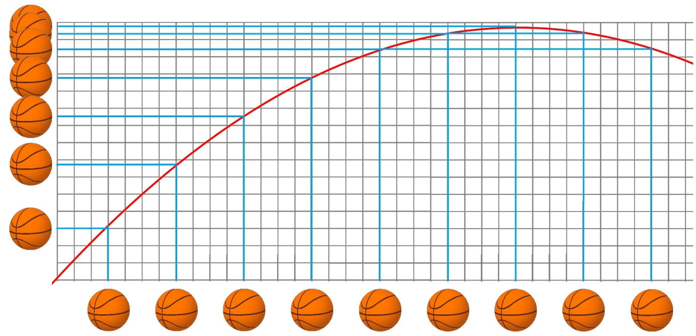
What if we do both: We roll a soccer ball horizontally and throw a ping-pong ball vertically? Let's try to follow both balls at the same time:



We'd have to fly through the air as if thrown at an angle!

Our understanding is that a thrown ball moves in both vertical and horizontal directions, simultaneously and independently:





The dynamics is very different:

1. In the horizontal direction, as there is no force changing the velocity, the latter remains constant.
2. Meanwhile, the vertical velocity is constantly changed by the force of gravity.

Let's now use these descriptions to represent the motion mathematically.

Recall how we used these *difference quotients* to find velocity and then the acceleration from the location:

$$\boxed{p_n \quad \mapsto \quad v_n = \frac{\Delta p}{\Delta t} = \frac{p_{n+1} - p_n}{h} \quad \mapsto \quad a_n = \frac{\Delta v}{\Delta t} = \frac{v_{n+1} - v_n}{h}}$$

Here $h = \Delta t$ is the increment of time.

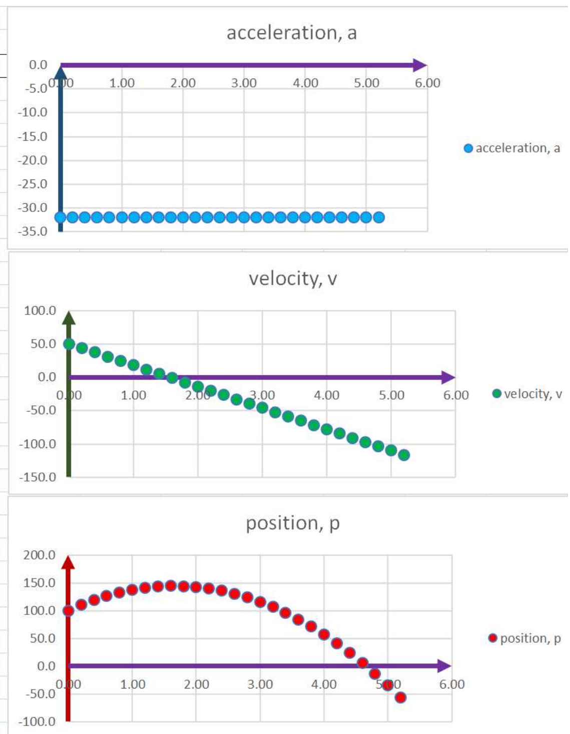
These formulas have also been solved for p_{n+1} and v_{n+1} respectively in order to model location as a function of time:

$$\boxed{a_n \quad \mapsto \quad v_{n+1} = v_n + ha_n \quad \mapsto \quad p_{n+1} = p_n + hv_n}$$

These recursive formulas are the *Riemann sums*: The displacements are being added one at a time.

This is what the results might look like:

h	t	a	v	p
0.20				
intial:	*0.00*	*-32.00*	*50.00*	*100.00*
	0.20	-32.00	43.60	110.00
	0.40	-32.00	37.20	118.72
	0.60	-32.00	30.80	126.16
	0.80	-32.00	24.40	132.32
	1.00	-32.00	18.00	137.20
	1.20	-32.00	11.60	140.80
	1.40	-32.00	5.20	143.12
	1.60	-32.00	-1.20	144.16
	1.80	-32.00	-7.60	143.92
	2.00	-32.00	-14.00	142.40
	2.20	-32.00	-20.40	139.60
	2.40	-32.00	-26.80	135.52
	2.60	-32.00	-33.20	130.16
	2.80	-32.00	-39.60	123.52
	3.00	-32.00	-46.00	115.60
	3.20	-32.00	-52.40	106.40
	3.40	-32.00	-58.80	95.92
	3.60	-32.00	-65.20	84.16
	3.80	-32.00	-71.60	71.12
	4.00	-32.00	-78.00	56.80
	4.20	-32.00	-84.40	41.20
	4.40	-32.00	-90.80	24.32
	4.60	-32.00	-97.20	6.16
	4.80	-32.00	-103.60	-13.28
	5.00	-32.00	-110.00	-34.00
	5.20	-32.00	-116.40	-56.00
	5.40	-32.00	-122.80	-79.28
	5.60	-32.00	-129.20	-103.84
	5.80	-32.00	-135.60	-129.68



This time, we have *two* such sequences: one for horizontal and one for vertical.

We construct the Cartesian coordinate system in the most convenient way:

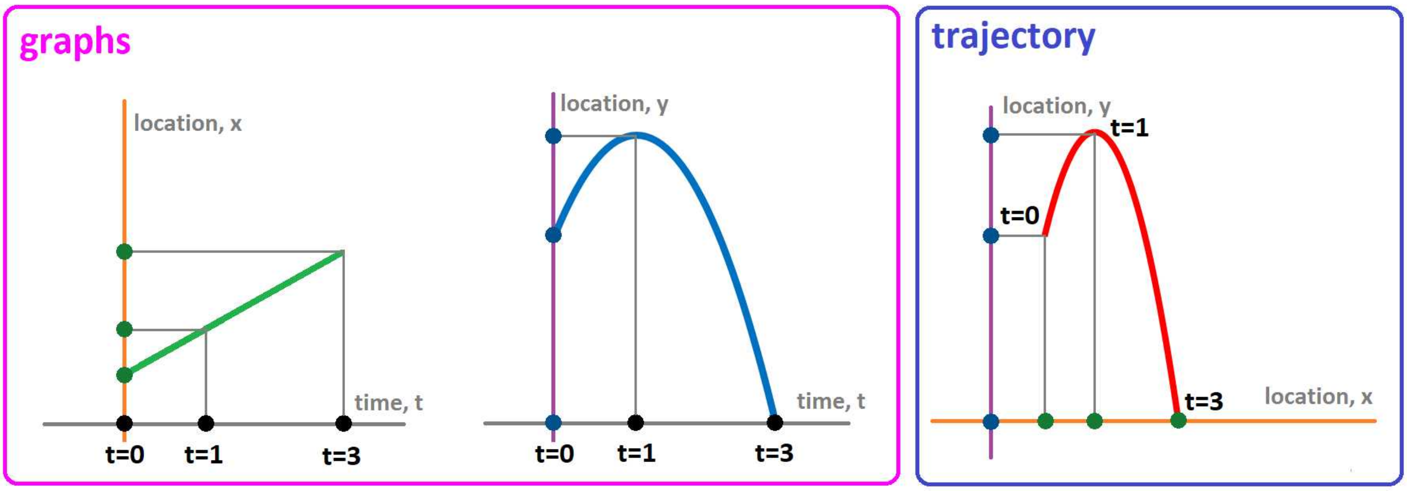
- The x-axis is horizontal.
- The y-axis is vertical.

However, we abandon the familiar $y = f(x)$ setup! We have *three variables* now:

1. t is time.
2. x is the horizontal dimension, the depth.
3. y is the vertical dimension, the height.

Either of the two *spatial* variables depends on the *temporal* variable.

Their *graphs* are plotted below (left):



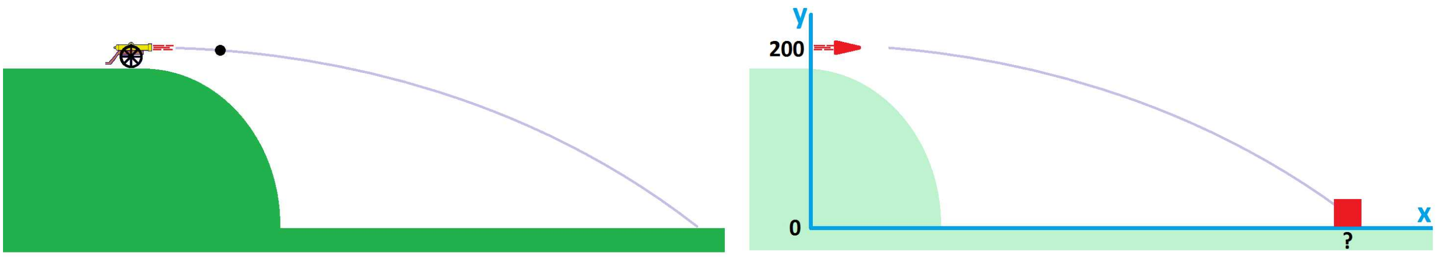
Meanwhile, the actual *path* of the ball through space will appear to an observer as a curve in the, vertically aligned, xy-plane (right). As there is no t-axis, we provide the times by labeling a few points of the trajectory.

Historically, one of the very first applications of calculus was in *ballistics*.

Before calculus, one had to resort to trial and error and watching where the cannonballs were landing. A well-designed test may provide one with a table (i.e., a function) that gives the shot length for each angle of the barrel. However, such a reference table may prove useless when one is to shoot from an elevated position, or at an elevated target, or over an obstacle. There are just too many parameters!

Let's consider one of such problems.

> ▶ *PROBLEM:* From a 200-foot elevation, a cannon is fired horizontally at 200 feet per second. How far will the cannonball go?



We will find the whole path!

Let $h = \Delta t$ be the increment of time. We have these *six* sequences with the difference quotients computed four times:

	horizontal	vertical	
position	x_n	y_n	
velocity	$v_n = \dfrac{x_{n+1} - x_n}{h}$	$u_n = \dfrac{y_{n+1} - y_n}{h}$	DQ
acceleration	$a_n = \dfrac{v_{n+1} - v_n}{h}$	$b_n = \dfrac{u_{n+1} - u_n}{h}$	DQ

Now, from the purpose of modeling and simulation, the derivation should go in the opposite direction. We go in reverse: the velocity and then the location from the acceleration. When we solve the above equations, we end up with these *four* recursive formulas (the Riemann sums) for our six sequences:

	horizontal	vertical	
acceleration	a_n	b_n	
velocity	$v_{n+1} = v_n + ha_n$	$u_{n+1} = u_n + hb_n$	RS
position	$x_{n+1} = x_n + hv_n$	$y_{n+1} = y_n + hu_n$	RS

Example 2.13.1: how far

Now in the specific case of *free fall*, there is just one force, the gravity. Therefore, the horizontal acceleration is zero and the vertical acceleration is constant (feet per second squared):

$$a = 0,\ b = -32\,.$$

We choose the increment of time:

$$h = 0.1\,.$$

Next, we acquire the initial conditions:

	x	y
initial location:	$x_0 = 0$	$y_0 = 200$
initial velocity:	$v_0 = 200$	$u_0 = 0$

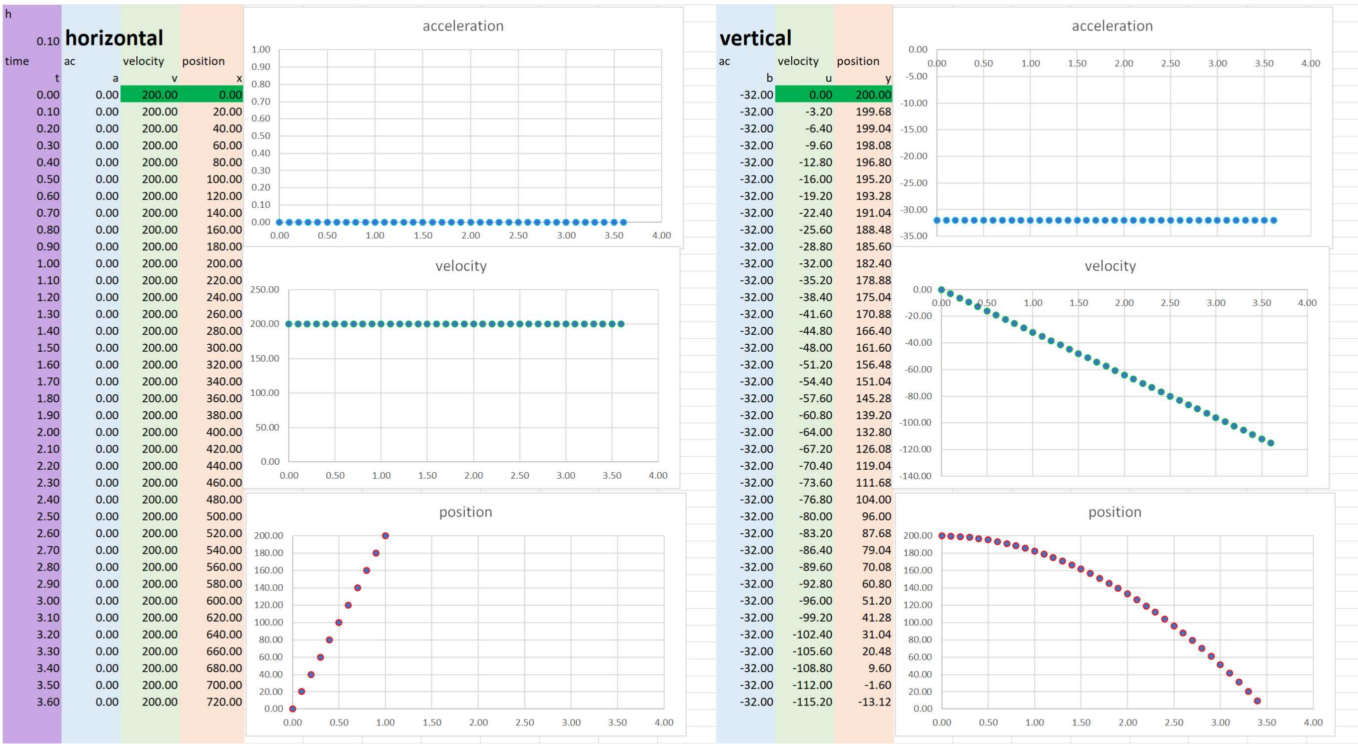
These four numbers serve as the *initial terms* of our four sequences:

time		horizontal	vertical
t_0	acceleration	$a_0 = 0$	$b_0 = -32$
	velocity	$v_1 = 200 + 0.1 \cdot 0$	$u_1 = 0 + 0.1 \cdot (-32)$
	position	$x_1 = 0 + 0.1 \cdot 200$	$y_1 = 200 + 0.1 \cdot 0$
$t_1 = t_0 + h$	acceleration	$a_1 = 0$	$b_1 = -32$
	velocity	$v_2 = v_1 + ha_1$	$u_2 = u_1 + hb_1$
	position	$x_2 = x_1 + hv_1$	$y_3 = y_1n + hu_1$
$t_2 = t_1 + h$	...	...	...

The four formulas are identical just as before:

```
=R[-1]C+RC[-1]*R1C1
```

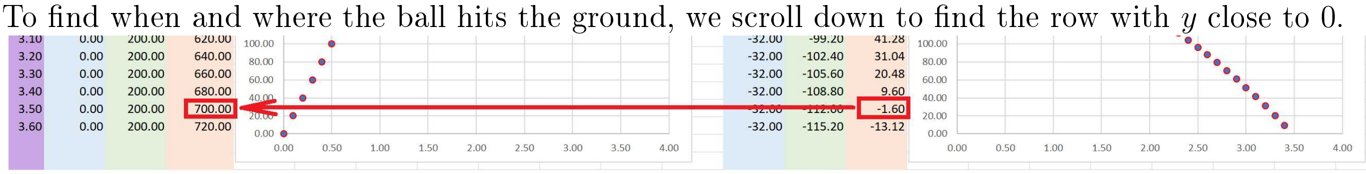
We use them to evaluate the location every $h = 0.1$ second. We take the spreadsheet presented above, copy and paste the columns for acceleration, velocity, and position:



time	ac	velocity	position	ac	velocity	position
t	a	v	x	b	u	y
0.00	0.00	200.00	0.00	-32.00	0.00	200.00
0.10	0.00	200.00	20.00	-32.00	-3.20	199.68
0.20	0.00	200.00	40.00	-32.00	-6.40	199.04
0.30	0.00	200.00	60.00	-32.00	-9.60	198.08
0.40	0.00	200.00	80.00	-32.00	-12.80	196.80
0.50	0.00	200.00	100.00	-32.00	-16.00	195.20
0.60	0.00	200.00	120.00	-32.00	-19.20	193.28
0.70	0.00	200.00	140.00	-32.00	-22.40	191.04
0.80	0.00	200.00	160.00	-32.00	-25.60	188.48
0.90	0.00	200.00	180.00	-32.00	-28.80	185.60
1.00	0.00	200.00	200.00	-32.00	-32.00	182.40
1.10	0.00	200.00	220.00	-32.00	-35.20	178.88
1.20	0.00	200.00	240.00	-32.00	-38.40	175.04
1.30	0.00	200.00	260.00	-32.00	-41.60	170.88
1.40	0.00	200.00	280.00	-32.00	-44.80	166.40
1.50	0.00	200.00	300.00	-32.00	-48.00	161.60
1.60	0.00	200.00	320.00	-32.00	-51.20	156.48
1.70	0.00	200.00	340.00	-32.00	-54.40	151.04
1.80	0.00	200.00	360.00	-32.00	-57.60	145.28
1.90	0.00	200.00	380.00	-32.00	-60.80	139.20
2.00	0.00	200.00	400.00	-32.00	-64.00	132.80
2.10	0.00	200.00	420.00	-32.00	-67.20	126.08
2.20	0.00	200.00	440.00	-32.00	-70.40	119.04
2.30	0.00	200.00	460.00	-32.00	-73.60	111.68
2.40	0.00	200.00	480.00	-32.00	-76.80	104.00
2.50	0.00	200.00	500.00	-32.00	-80.00	96.00
2.60	0.00	200.00	520.00	-32.00	-83.20	87.68
2.70	0.00	200.00	540.00	-32.00	-86.40	79.04
2.80	0.00	200.00	560.00	-32.00	-89.60	70.08
2.90	0.00	200.00	580.00	-32.00	-92.80	60.80
3.00	0.00	200.00	600.00	-32.00	-96.00	51.20
3.10	0.00	200.00	620.00	-32.00	-99.20	41.28
3.20	0.00	200.00	640.00	-32.00	-102.40	31.04
3.30	0.00	200.00	660.00	-32.00	-105.60	20.48
3.40	0.00	200.00	680.00	-32.00	-108.80	9.60
3.50	0.00	200.00	700.00	-32.00	-112.00	-1.60
3.60	0.00	200.00	720.00	-32.00	-115.20	-13.12

Of course, for the horizontal values, we replace acceleration with $a = 0$.

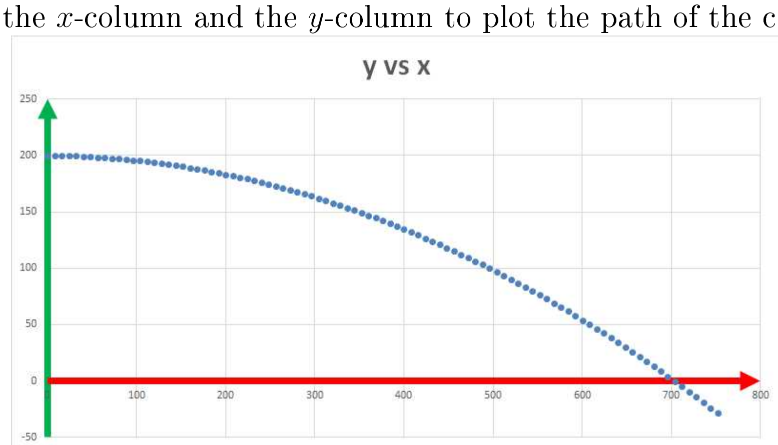
To find when and where the ball hits the ground, we scroll down to find the row with y close to 0.



3.10	0.00	200.00	620.00	-32.00	-99.20	41.28
3.20	0.00	200.00	640.00	-32.00	-102.40	31.04
3.30	0.00	200.00	660.00	-32.00	-105.60	20.48
3.40	0.00	200.00	680.00	-32.00	-108.80	9.60
3.50	0.00	200.00	700.00	-32.00	-112.00	-1.60
3.60	0.00	200.00	720.00	-32.00	-115.20	-13.12

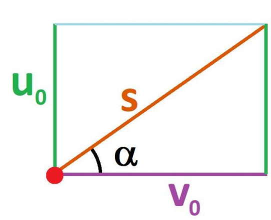
It happens about $t = 3.5$. Then, the value of x at the time is about $x = 700$.

We can also combine the x-column and the y-column to plot the path of the cannonball:



With the spreadsheet, we can ask and answer a variety of questions about such motion. But first, let's introduce the *angle* of the barrel of the cannon into the model.

The velocity of 200 feet per second we have been using is the "muzzle velocity", i.e., the *speed*, s, with which the cannonball leaves the muzzle – no matter what the angle, α, is. That's where the *initial* horizontal and the vertical velocities come from:

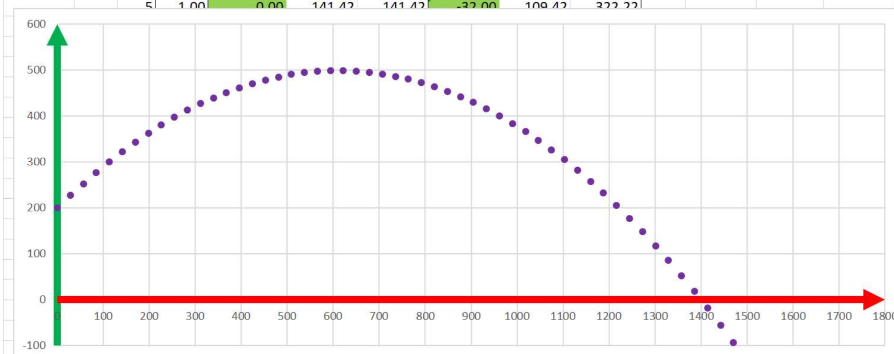


These formulas come from trigonometry (Volume 1):

$$v_0 = s\cos\alpha \quad \text{and} \quad u_0 = s\sin\alpha\,.$$

We use them below to provide the initial values of the two velocities:

h=	0.20			speed s =	200.00				
				angle =	0.79				
		n	t	a	v	x	b	u	y
	initial:	0	0.00	0.00	141.42	0.00	-32.00	141.42	200.00
		1	0.20	0.00	141.42	28.28	-32.00	135.02	227.00
		2	0.40	0.00	141.42	56.57	-32.00	128.62	252.73
		3	0.60	0.00	141.42	84.85	-32.00	122.22	277.17
		4	0.80	0.00	141.42	113.14	-32.00	115.82	300.34
		5	1.00	0.00	141.42	141.42	-32.00	109.42	322.22



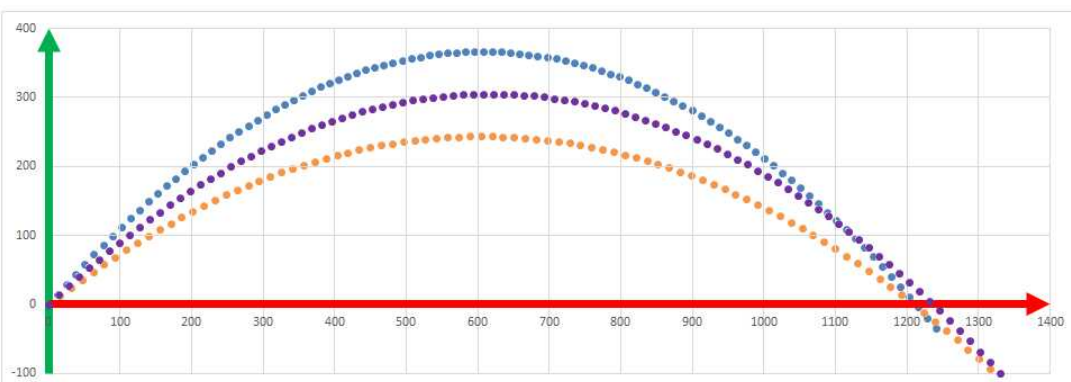
We can freely enter the data for the following (highlighted in green):

- the initial speed
- the initial angle
- the initial location
- all accelerations

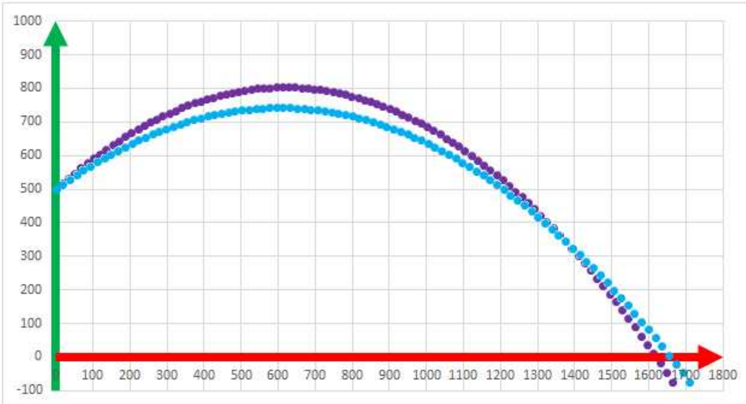
The rest is computed according to the same formulas as before.

Example 2.13.2: longest shot

It is really true that 45 degrees is the best angle to shoot for a longer distance:



It appears that the one in the middle is the best, but we can't prove this with just the numerical methods. Now, what if we try to shoot from a hill again, say, 500 feet high?



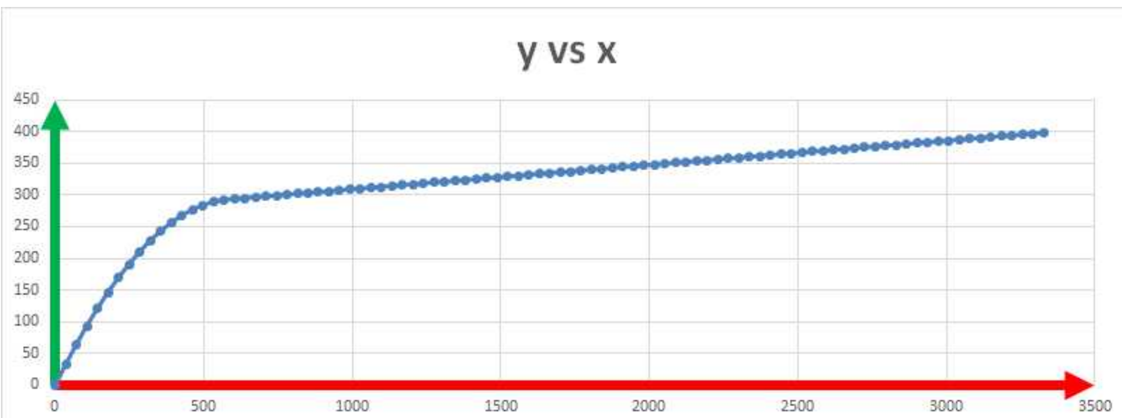
It's not the best anymore!

Exercise 2.13.3

Show that the best shot will become more and more flat as the elevation grows.

Example 2.13.4: variable gravity

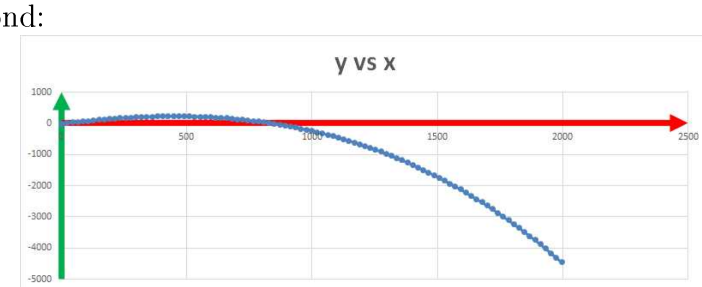
What happens if the gravity suddenly disappears? In the column for the vertical acceleration, we just replace -32 with 0 after a few rows:

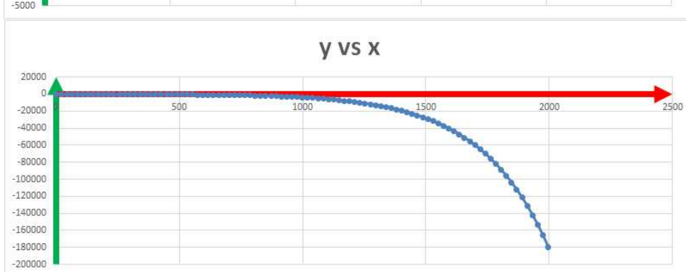


The cannonball flies off on a *tangent*.

Example 2.13.5: variable gravity

What happens if the gravity starts to increase? Let's increase the downward acceleration 1 foot per second squared per second:

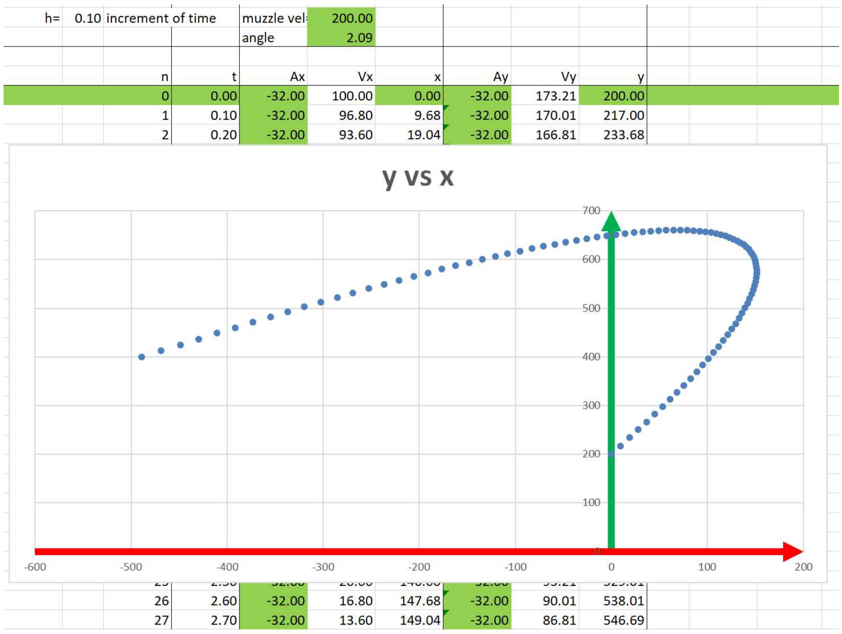




The trajectory looks steeper and steeper, but is there a *vertical asymptote*? We can't answer with just the numerical methods.

Example 2.13.6: horizontal gravity

What happens if the gravity is horizontal instead? The motion will be along a parabola that lies on its side, of course. But what if there are both vertical (down) and horizontal (left) forces of gravity? Let's modify the acceleration columns accordingly by replacing 0's with -32 in the horizontal acceleration column:



h= 0.10 increment of time; muzzle vel 200.00; angle 2.09

n	t	Ax	Vx	x	Ay	Vy	y
0	0.00	-32.00	100.00	0.00	-32.00	173.21	200.00
1	0.10	-32.00	96.80	9.68	-32.00	170.01	217.00
2	0.20	-32.00	93.60	19.04	-32.00	166.81	233.68
26	2.60	-32.00	16.80	147.68	-32.00	90.01	538.01
27	2.70	-32.00	13.60	149.04	-32.00	86.81	546.69

Is this a parabola?

Exercise 2.13.7

Explain the results in the last example.

In spite of these numerous examples, we can only do one at a time! The conclusions we draw are also specific to these situations (accelerations, initial conditions, etc.).

Example 2.13.8: explicit formulas

Because everything is recursive, we have to run the computations directly to see what happens, as in the last example:

acceleration	a_n
velocity	$v_{n+1} = v_n + ha_n$
position	$x_{n+1} = x_n + hv_n$

Is there an explicit formula for the position? In other words, can we express x_n in terms of n?

If the acceleration is zero, it's easy:

$$a_n = 0 \implies v_n = v_0 = v \implies x_{n+1} = x_n + hv\,.$$

Adding the same number is just multiplication:

$$x_n = x_0 + hvn\,.$$

If the acceleration is non-zero but constant, it's more complicated:

$$a_n = a \implies v_{n+1} = v_n + ha\,.$$

Therefore, we have:

$$v_n = v_0 + han\,.$$

Next:

$$x_{n+1} = x_n + hv_n = x_n + h(v_0 + han) = x_n + hv_0 + h^2an\,.$$

We are adding consecutive integers!

Exercise 2.13.9

Finish the computation. Hint: Use a formula from Chapter 1.

Exercise 2.13.10

What if the acceleration is increasing linearly? Hint: Use a formula from Chapter 1.

The challenge of finding explicit formulas for sums of sequences is addressed in Chapter 5.

This is why we now turn to the *continuous* case, i.e., we take the limit of everything above:

$$\boxed{h = \Delta t \to 0}$$

The disappearance of h makes algebra simpler!

This time, instead of six *sequences*, we have these six *functions* of time:

x, the depth, the horizontal location	$v = x'$, the horizontal velocity	$a = v'$, the horizontal acceleration
y, the height, the vertical location	$u = y'$, the vertical velocity	$b = u'$, the vertical acceleration

There is no time increment as a parameter anymore!

Now the specific case of free fall:

$$a = 0,\ b = -g\,.$$

From Volume 2, we know:

1. The derivative of a quadratic polynomial is linear. And the only function the derivative of which is linear is a quadratic polynomial.
2. The derivative of a linear polynomial is constant. And the only function the derivative of which is constant is a linear polynomial.

We conclude about *free fall*:

1. The horizontal position $x = x(t)$ is linear.
2. The vertical position $y = y(t)$ is quadratic.

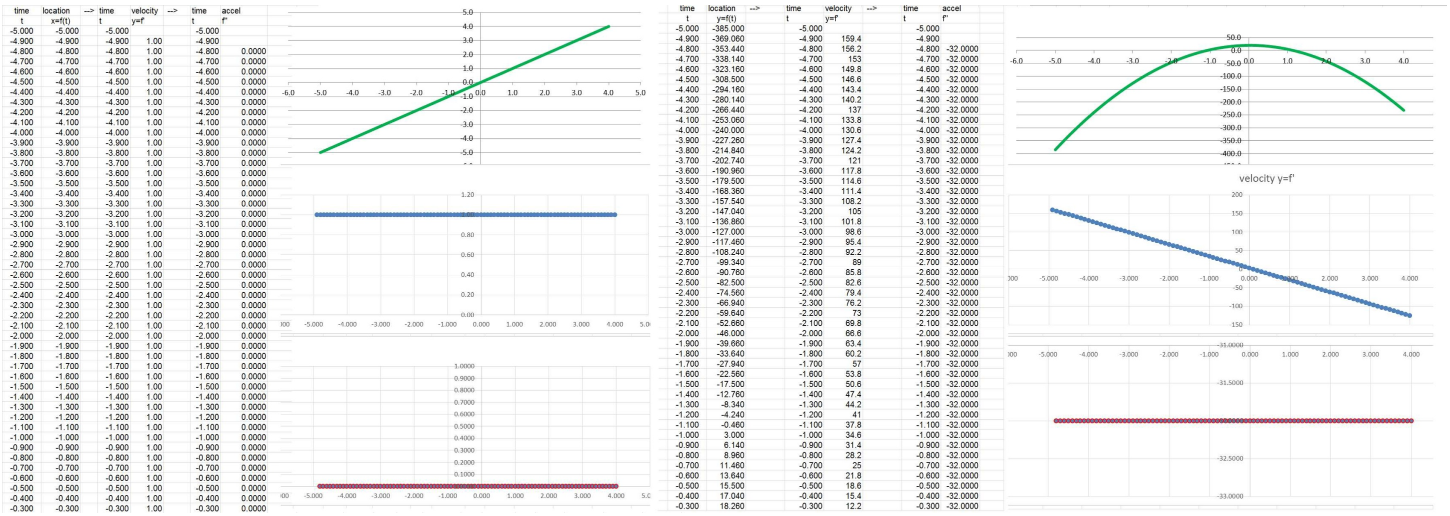
What makes these two functions specific are the *initial conditions*:

x_0, the initial depth, $x_0 = x(0)$	v_0, the initial horizontal component of velocity, $v(0) = \left.\dfrac{dx}{dt}\right\vert_{t=0}$
y_0, the initial height, $y_0 = y(0)$	u_0, the initial vertical component of velocity, $u(0) = \left.\dfrac{dy}{dt}\right\vert_{t=0}$

Therefore, we have:

$$\begin{array}{lllll} x & = x_0 & +v_0t & \\ y & = y_0 & +u_0t & -\dfrac{1}{2}gt^2 \end{array}$$

These two equations allow us to solve a variety of problems about motion. We carry this out for x and y separately and the results are shown in the spreadsheet:



Example 2.13.11: how far

Let's revisit the problem about a specific shot we solved numerically. Our equations become:

$$\begin{aligned} x &= \qquad 200t \\ y &= 200 \qquad\qquad -16t^2 \end{aligned}$$

Now, analytically, the height at the end is 0, so to find *when* it happened, we set $y = 0$, or

$$200 - 16t^2 = 0\,,$$

and solve for t. Then, the time of landing is:

$$t_1 = \sqrt{\frac{200}{16}} = \frac{5\sqrt{2}}{2}\,.$$

To find *where* it happened, we substitute this value of t into x; the location is:

$$x_1 = 200t_1 = 200\frac{5\sqrt{2}}{2} \approx 707\,.$$

The result matches our estimate!

Example 2.13.12: variable gravity

What happens if the gravity is decreasing? Suppose it is decreasing 1 foot per second squared per second. With the tools developed in this chapter, it's easy:

$$b = -g - t \implies u = u_0 - gt - \frac{t^2}{2} \implies y = y_0 + u_0 t - \frac{gt^2}{2} - \frac{t^3}{6}\,.$$

We confirm that the trajectory will become steeper and steeper. We also discover that there is no vertical asymptote!

Chapter 3: What we can do with integral calculus

Contents

3.1. The area between two graphs

There are two main ways to enter *differential* calculus: through the study of geometry (finding secant and tangent lines of curves) and through the study of motion (finding velocity and acceleration from location).

Similarly, there are two main ways to enter *integral* calculus: through the study of geometry – finding areas under curves – and through the study of motion – finding location from velocity and acceleration. These are two very distinct examples of *recognizing Riemann sums*.

Throughout this chapter, we will carry out this key step in a variety of entirely new situations. But we will start with something familiar.

Example 3.1.1: area of circle

In Chapter 1, we confirmed that the area of a circle of radius r is $A = \pi r^2$ using nothing but a spreadsheet. And later in the chapter, we used integration to provide a precise answer. The solution, however, wasn't fully satisfactory because we relied on the symmetry of the circle to compute the area of its half so that the area of the whole circle is then twice this number. This is too limiting. Let's start over.

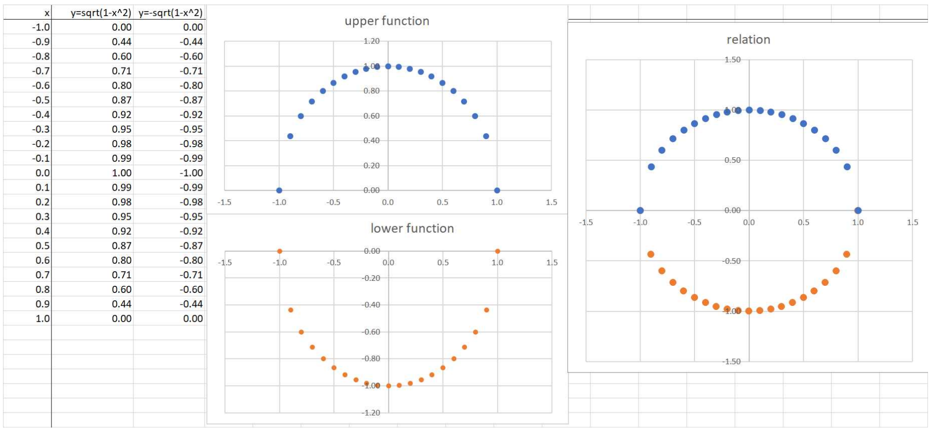
There are *two* functions this time, for the top and the bottom of the circle:

$$f(x) = \sqrt{1-x^2} \quad \text{and} \quad g(x) = -\sqrt{1-x^2}, \quad -1 \le x \le 1\,.$$

With the formulas:

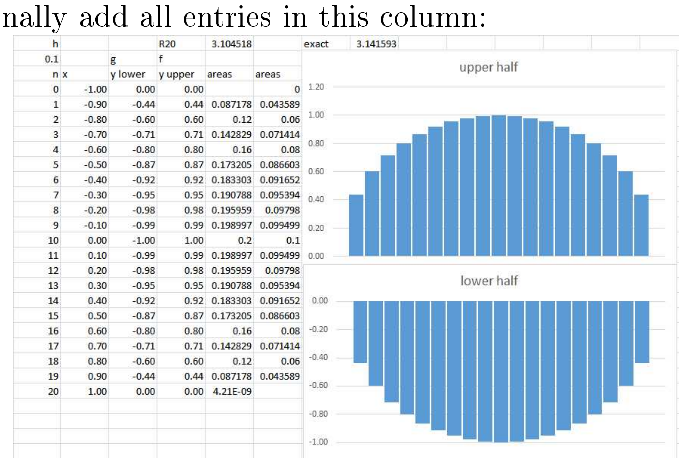
`=SQRT(1-RC[-2]^2)` and `=-SQRT(1-RC[-2]^2)`

we plot both:



x	y=sqrt(1-x^2)	y=-sqrt(1-x^2)
-1.0	0.00	0.00
-0.9	0.44	-0.44
-0.8	0.60	-0.60
-0.7	0.71	-0.71
-0.6	0.80	-0.80
-0.5	0.87	-0.87
-0.4	0.92	-0.92
-0.3	0.95	-0.95
-0.2	0.98	-0.98
-0.1	0.99	-0.99
0.0	1.00	-1.00
0.1	0.99	-0.99
0.2	0.98	-0.98
0.3	0.95	-0.95
0.4	0.92	-0.92
0.5	0.87	-0.87
0.6	0.80	-0.80
0.7	0.71	-0.71
0.8	0.60	-0.60
0.9	0.44	-0.44
1.0	0.00	0.00

We let the values of x run from -1 to 1 every 0.1 and covered, best we can, this circle with vertical bars based on these segments. Then the area of the circle is approximated by the sum of the areas of the bars: We add a column of the widths of the bars, multiply them by the heights, place the result in the last column, and finally add all entries in this column:



h			R20	3.104518		exact	3.141593
0.1		g	f				
n	x	y lower	y upper	areas	areas		
0	-1.00	0.00	0.00		0		
1	-0.90	-0.44	0.44	0.087178	0.043589		
2	-0.80	-0.60	0.60	0.12	0.06		
3	-0.70	-0.71	0.71	0.142829	0.071414		
4	-0.60	-0.80	0.80	0.16	0.08		
5	-0.50	-0.87	0.87	0.173205	0.086603		
6	-0.40	-0.92	0.92	0.183303	0.091652		
7	-0.30	-0.95	0.95	0.190788	0.095394		
8	-0.20	-0.98	0.98	0.195959	0.09798		
9	-0.10	-0.99	0.99	0.198997	0.099499		
10	0.00	-1.00	1.00	0.2	0.1		
11	0.10	-0.99	0.99	0.198997	0.099499		
12	0.20	-0.98	0.98	0.195959	0.09798		
13	0.30	-0.95	0.95	0.190788	0.095394		
14	0.40	-0.92	0.92	0.183303	0.091652		
15	0.50	-0.87	0.87	0.173205	0.086603		
16	0.60	-0.80	0.80	0.16	0.08		
17	0.70	-0.71	0.71	0.142829	0.071414		
18	0.80	-0.60	0.60	0.12	0.06		
19	0.90	-0.44	0.44	0.087178	0.043589		
20	1.00	0.00	0.00	4.21E-09			

The height of the rectangle located at x is $f(x)-g(x)$, and its area is $(f(x)-g(x))\cdot 0.1$. We compute these in the next column and then add them:

$$\text{approximate area of the circle} = 3.1\,.$$

It is close to the theoretical result we established in Chapter 1:

$$\text{exact area of the circle} = \pi = 3.14159...$$

Of course, we realize that we could produce the same result if we take the data from the first spreadsheet,

$$\sum_i f(c_i)\cdot 0.1\,,$$

and then subtract the data for the new function,

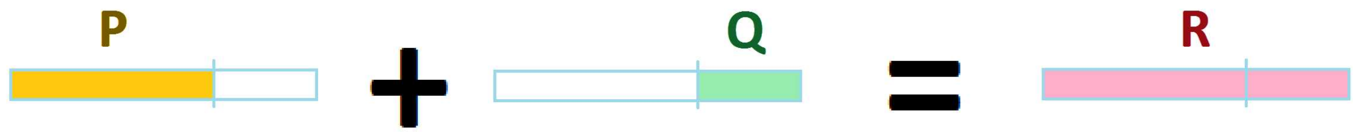
$$\sum_i g(c_i)\cdot 0.1\,.$$

Furthermore, we have

$$\sum_i f(c_i)0.1 - \sum_i g(c_i)\cdot 0.1 = \sum_i \big(f(c_i)-g(c_i)\big)\cdot 0.1\,.$$

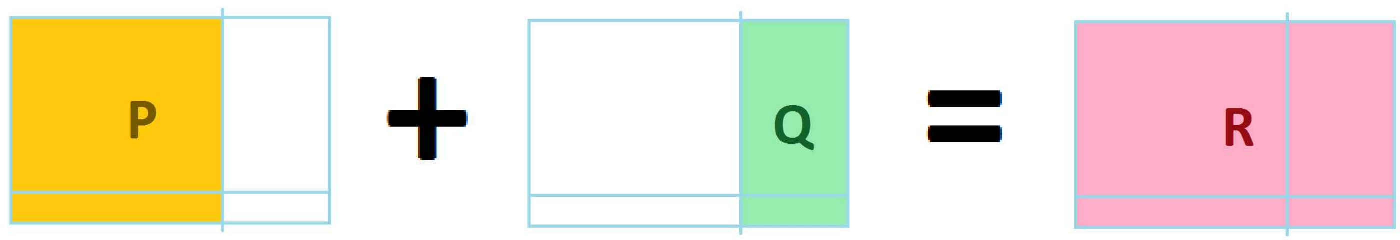
The common sense about how the (unsigned) lengths of intervals behave is that the length of the union of two intervals is the sum of the two lengths minus the lengths of the intersection:

$$\text{length of } P\cup Q = \text{ length of } P + \text{ length of } Q - \text{ length of } P\cap Q\,.$$



It is called the *additivity of the length*. The last term disappears when there is no overlap or it is just a point.

If we build rectangles on top of these intervals, we are in a similar situation – for the (unsigned) areas:

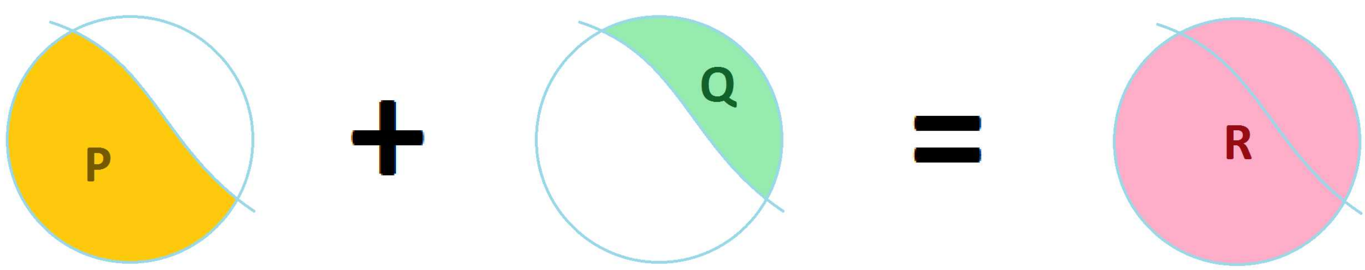


In other words, the area of the union of two regions is the sum of the two areas minus the area of the intersection:

$$\text{area of } P\cup Q = \text{ area of } P + \text{ area of } Q - \text{ area of } P\cap Q\,.$$

It is called the *additivity of the area*. The last term disappears when there is no overlap or it is just a curve.

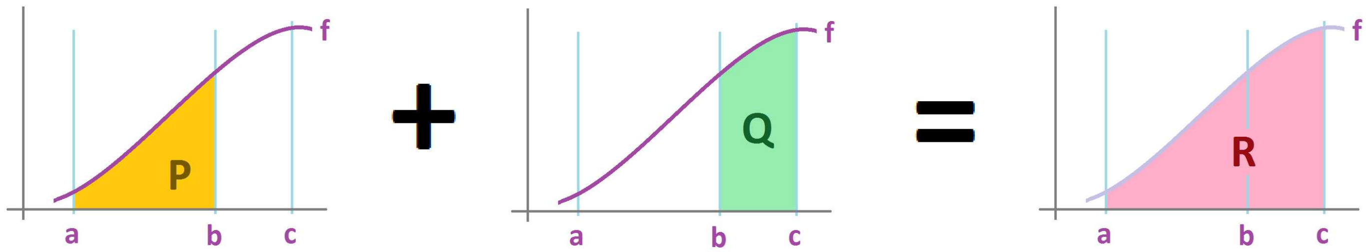
The idea is then to be applied to curved regions:



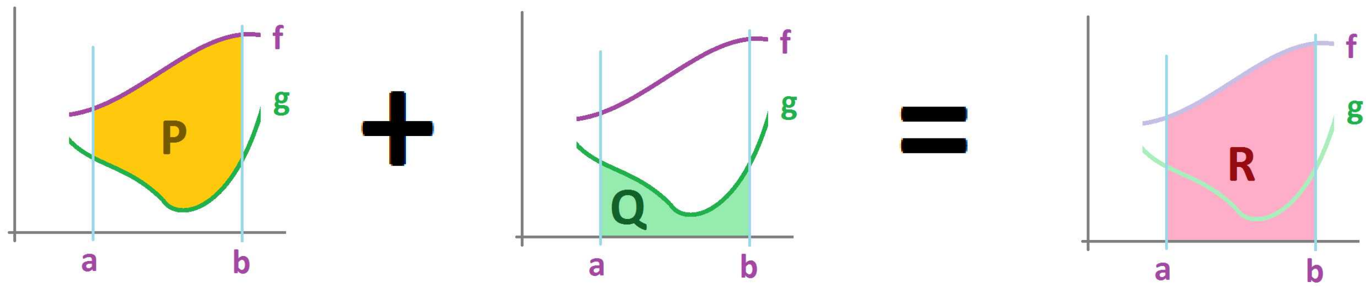
However, our understanding of areas is limited to those of *regions under graphs* of functions. Even then, the additivity of the areas of those regions has been only demonstrated for the special case when such a region is cut by a vertical line:

$$\int_a^b f\,dx + \int_b^c f\,dx = \int_a^c f\,dx\,.$$

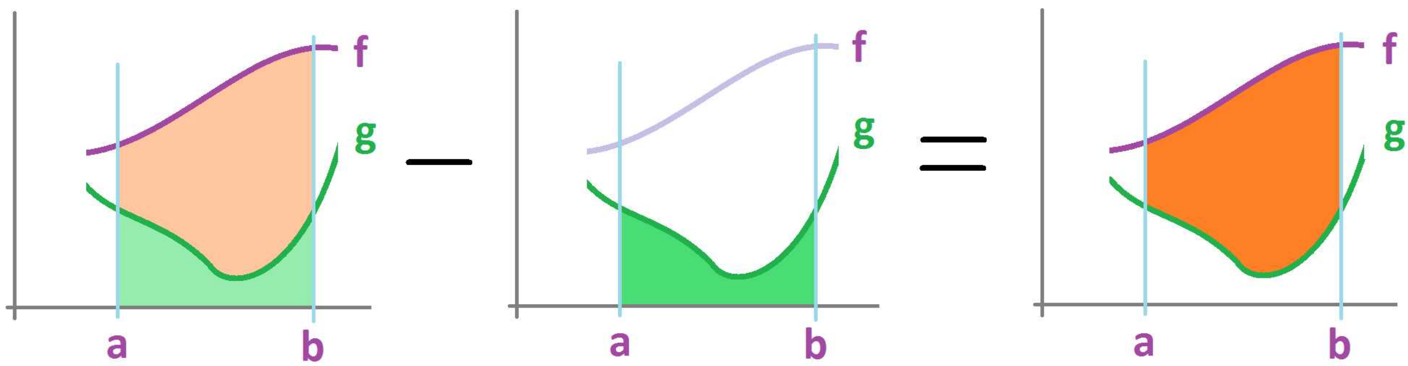
This case is illustrated below (the line is $x=b$):



What if the region under the graph is cut by another graph?



We would like to find the area of the region *between the graphs*:



The integral interpretation is easy to see: If $f(x) \geq g(x)$ for all x in $[a,b]$, then:

$$P = R - Q = \int_a^b f\,dx - \int_a^b g\,dx = \int_a^b (f-g)\,dx\,.$$

We have assumed the additivity of the areas and used the *Sum Rule for integrals*.

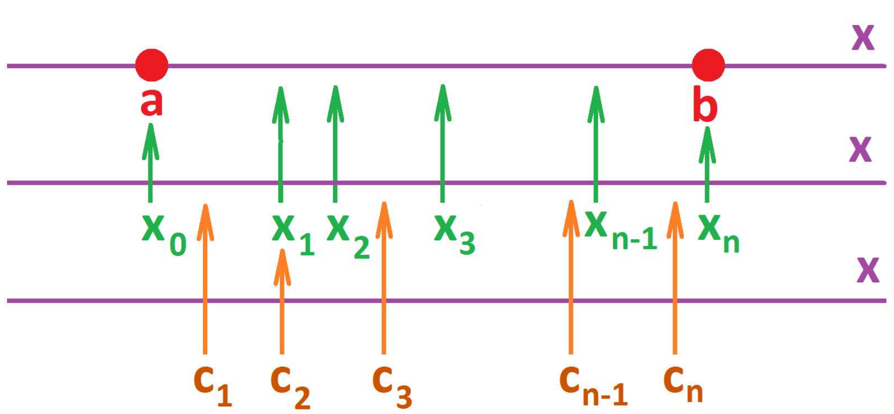
However, every term in the formula is the area *under the graph*. In order to justify the additivity for the areas *between the graphs*, we need to start from scratch.

Back to approximations.

We start, as before, with a partition of the interval $[a,b]$ into n intervals of possibly different lengths:

$$[x_0, x_1],\ [x_1, x_2],\ \ldots,\ [x_{n-1}, x_n]\,,$$

with $x_0 = a,\ x_n = b$.



The primary nodes of P are:

$$x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n$$

The lengths of the intervals are:

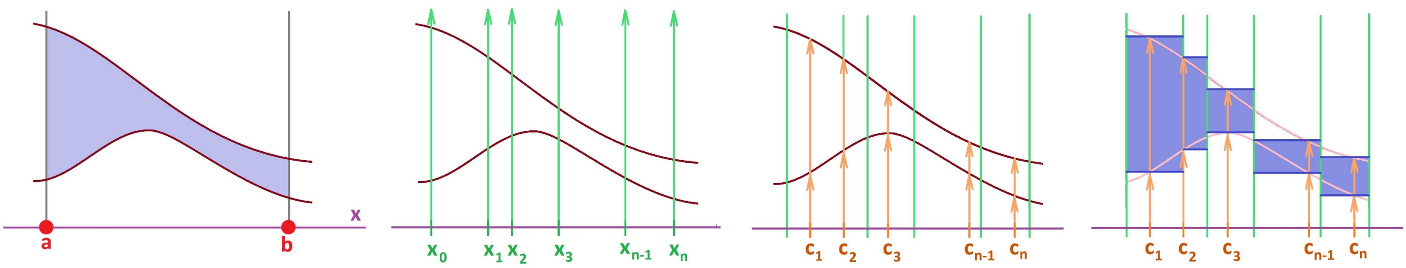
$$\Delta x_i = x_i - x_{i-1},\ i = 1, 2, \ldots, n\,.$$

The secondary nodes of P are:

$$c_1 \text{ in } [x_0, x_1],\ c_2 \text{ in } [x_1, x_2],\ \ldots,\ c_n \text{ in } [x_{n-1}, x_n]\,.$$

This time, we face *two* functions.

We approximate the area between the graphs with rectangles with these widths:



Let's take a look at the ith rectangle. Its width is, as before, Δx_i. Now, its top is $f(c_i)$ and the bottom is $g(c_i)$ (instead of the x-axis). Therefore, its height is $f(c_i) - g(c_i)$. Then, its area is $(f(c_i) - g(c_i))\Delta x_2$. Hence, the total area of the rectangles is:

$$(f(c_1) - g(c_1))\Delta x_1 + (f(c_2) - g(c_2))\Delta x_2 + ... + (f(c_n) - g(c_n))\Delta x_n \,.$$

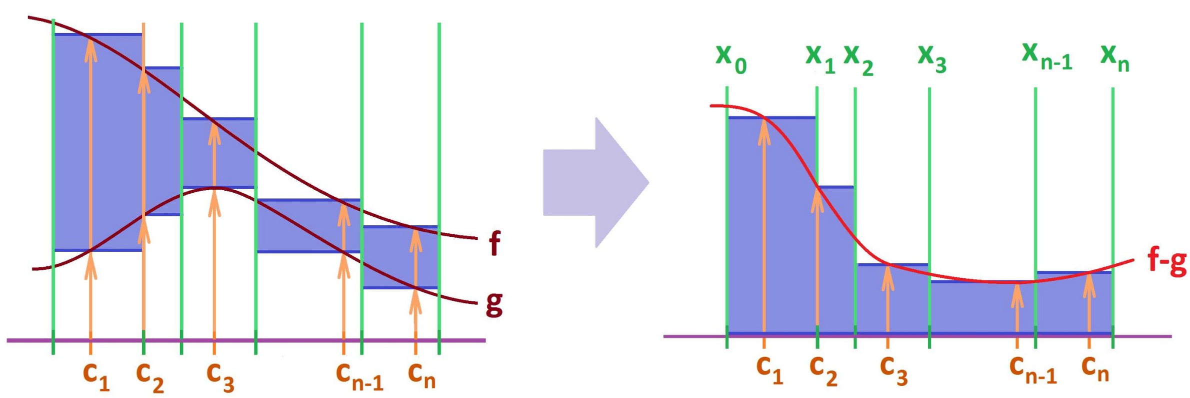
The key step is the following. We *recognize* this expression as the Riemann sum of this new function, $f - g$, the *difference*:

$$\Sigma\,(f-g)\cdot\Delta x = \underbrace{\sum_{i=1}^{n}(f-g)(c_i)\Delta x_i}_{\text{areas of the rectangles}}\,.$$

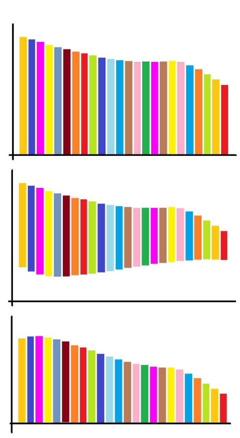
Indeed, it is convenient to think of each term as if it refers to a single function, $f - g$:

$$f(c_k) - g(c_k) = (f-g)(c_k)\,.$$

The rectangles we started with are shown on the left and the Riemann sum of the difference on the right:



We can still go back and explain the Riemann sum of the new function in terms of areas. It's as if the rectangles are first aligned with $y = f(x)$, then cut from below with $y = g(x)$, suspended in the air, and then dropped on the x-axis, like this:



What we see is the area under the graph of $f - g$.

We define the area accordingly:

Definition 3.1.2: area between graphs

Suppose f and g are two functions defined on interval $[a, b]$ with $f(x) \geq g(x)$ for all x in $[a, b]$. Then the *area between the graphs* of these functions over interval $[a, b]$ is defined to be the limit of a sequence of the Riemann sums of their difference with the mesh of their augmented partitions P_k approaching 0 as $k \to \infty$, when all these limits exist and are all equal to each other:

$$\text{Area between the graphs of } f, g \;= \lim_{k \to \infty} \Sigma \, (f - g) \cdot \Delta x \,.$$

Warning!

Unlike the area "under" the graph, this number cannot be negative as defined.

Exercise 3.1.3

If f and g represent the velocities of two objects, what does the area represent?

The definition repeats that of the Riemann integral:

Theorem 3.1.4: Area Between Graphs as Integral

Suppose f and g are two functions defined on interval $[a, b]$ with $f(x) \geq g(x)$ for all x in $[a, b]$. If $f - g$ is integrable, then the area between the graphs of f and g is equal to

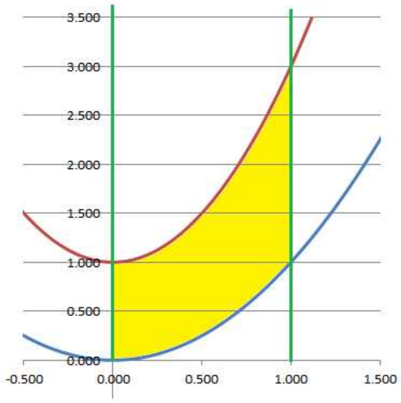
$$\boxed{\textit{Area between the graphs of } f, g \;= \int_a^b (f - g)\, dx}$$

We have a variety of regions we used to be unable to compute.

Example 3.1.5: area between parabolas

Evaluate the area of the region bounded by the parabolas $y = x^2$ and $y = 2x^2 + 1$ between $x = 0$ and $x = 1$.

It is clear that $g(x) = x^2$ and $f(x) = 2x^2 + 1$, as well as $a = 0$ and $b = 1$. The functions are continuous and, therefore, integrable. Before we apply the formula, we just need to confirm that the graph of f is above the graph of g:



For every x between 0 and 1, we have $x^2 < 2x^2+1$ because $0 < x^2+1$. Thus,

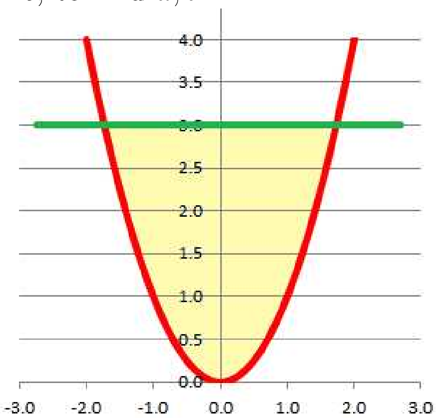
$$\text{Area } = \int_a^b (f-g)\,dx = \int_0^1 \left((2x^2+1)-x^2\right)\,dx = \int_0^1 (x^2+1)\,dx = \frac{1}{3}x^3 + x\Big|_0^1 = \frac{1}{3}+1 = \frac{4}{3}\,.$$

Sometimes the interval is not provided.

Example 3.1.6: area parabola and a line

Evaluate this area of the region is bounded by the parabola $y = x^2$ and the horizontal line $y = 3$.

We will need some algebra this time, to find a, b:



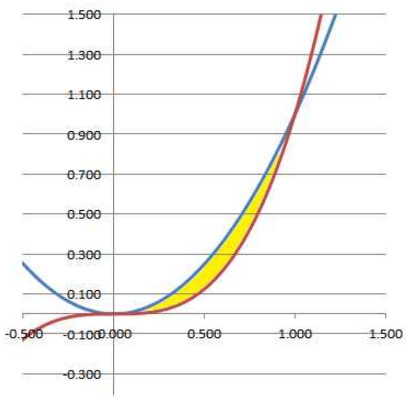
The intersection points (x, y) satisfy: $y = 3 = x^2$. Then $a = -\sqrt{3}$, $b = \sqrt{3}$. We also realize from the sketch that $f(x) = 3$ and $g(x) = x^2$. Then,

$$\begin{aligned}
\text{Area } &= \int_a^b (f-g)\,dx \\
&= \int_{-\sqrt{3}}^{\sqrt{3}} \left(3-x^2\right)\,dx \\
&= 3x - \frac{1}{3}x^3\Big|_{-\sqrt{3}}^{\sqrt{3}} \\
&= \left(3\sqrt{3} - \frac{1}{3}\sqrt{3}^3\right) - \left(-3\sqrt{3} - \frac{1}{3}(-\sqrt{3})^3\right) \\
&= 2\left(3\sqrt{3} - \frac{1}{3}\sqrt{3}^3\right) \\
&= 2\left(3\sqrt{3} - \sqrt{3}\right) \\
&= 4\sqrt{3}\,.
\end{aligned}$$

Example 3.1.7: area between x^2 and x^3

Evaluate the area between $y = x^2$ and $y = x^3$.

Once again, we find the intersection points by solving $x^2 = x^3$. We have $a = 0$ and $b = 1$, which confirms the sketch and the fact that $x^3 < x^2$:

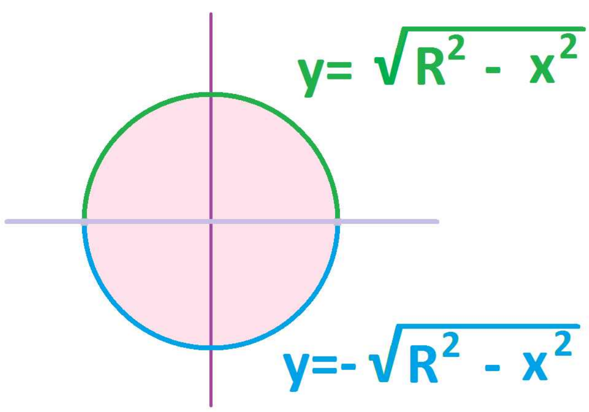


Then, we have:

$$\begin{aligned} \text{Area} &= \int_a^b (f-g)\,dx \\ &= \int_0^1 \left(x^2-x^3\right)\,dx \\ &= \frac{1}{3}x^3 - \frac{1}{4}x^4\bigg|_0^1 \\ &= \frac{1}{3}-\frac{1}{4} \\ &= \frac{1}{12}\,. \end{aligned}$$

Example 3.1.8: circle

Let's revisit the computation of the area of the circle:



This time we don't have to split it in half and rely on its symmetry; the circle is the region between two graphs:

$$y=\sqrt{R^2-x^2} \quad\text{and}\quad y=-\sqrt{R^2-x^2}\,.$$

The former is f and the latter is g. Also, $a=-R,\ b=R$. Then,

$$\text{Area} = \int_{-R}^{R}\left(\sqrt{R^2-x^2}+\sqrt{R^2-x^2}\right)\,dx = 2\int_{-R}^{R}\sqrt{R^2-x^2}\,dx = \pi R^2\,.$$

The integral is evaluated via a trig substitution, just as before.

Exercise 3.1.9

Find the area of the intersection of the two regions bounded by the circles $x^2+y^2=1$ and $(x-1)^2+y^2=1$.

Exercise 3.1.10

Find the area between the curves $x = y^2$ and $x = y^4$.

3.2. The linear density and the mass

The method that starts to shape up is as follows.

Suppose we have a quantity Q "contained" in a space region R: area, volume, mass, particular material, charge, etc. Then:

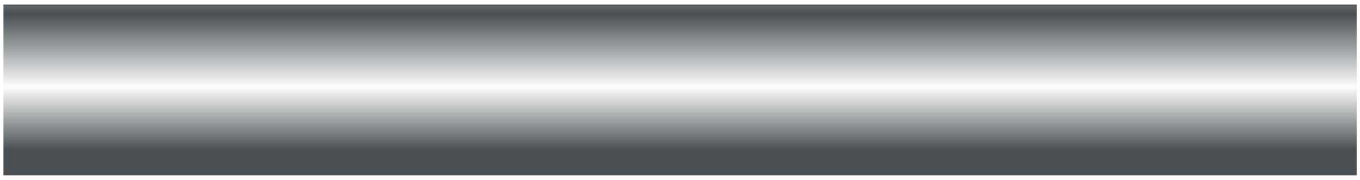
1. We represent the total quantity Q as the sum of its values Q_i over simpler, and smaller, parts of R.
2. We represent, or approximate, each of these values via a familiar quantity, e.g., area via length, volume via area, etc.
3. We recognize the sum as the Riemann sum of a function that represents some other quantity q spread over the region.
4. The quantity Q is equal to the integral of q.

The last step is necessary only when we approximate an idealized situation.

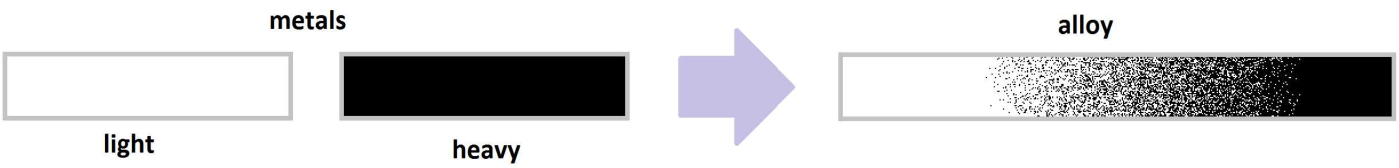
We will illustrate the method with one more example.

Let's recall how the *linear density* was defined in Volume 2.

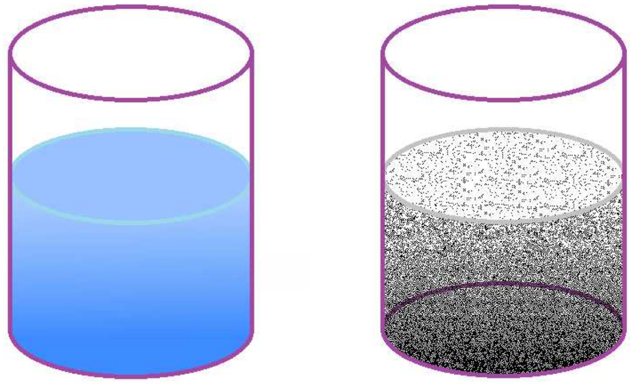
We are given a metal rod:



The rod might be non-uniform, i.e., the density varies but only in the horizontal direction. For example, this is what might happen when two metals are (imperfectly) melted into a piece of alloy:



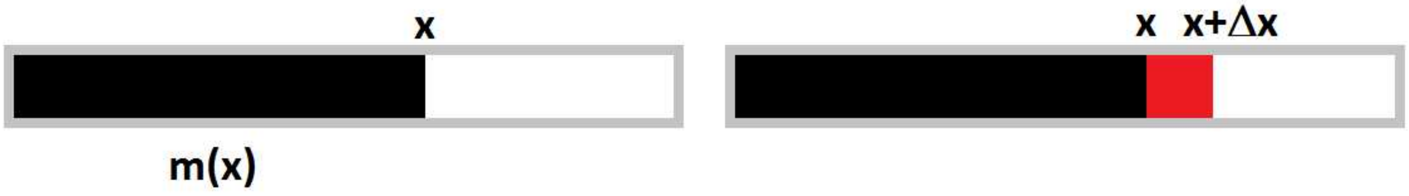
Another example is particles suspended in a liquid that settles – because of gravity – in a pattern that is denser at the bottom:



In either case, there is a line (we call it the x-axis) with no change in density in the directions perpendicular to it. We then ignore those directions and the density becomes a function of a single number x designating the location along this line; hence the *linear* density $y = l(x)$.

Take a small piece of the rod at location x, Δx long, and let's call its mass Δm. Then, for this piece, we have:

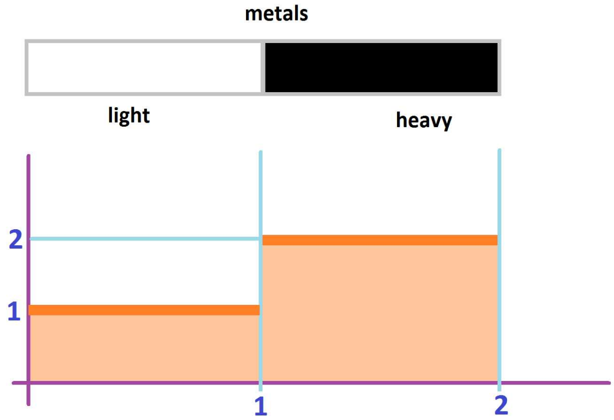
$$\text{Linear density} = \frac{\text{mass}}{\text{length}} = \frac{\Delta m}{\Delta x} = \frac{m(x+\Delta x) - m(x)}{\Delta x}\,.$$



Let's reverse this analysis. Suppose this time the linear density l is given, what is the mass of the rod?

Example 3.2.1: two pieces

Suppose the two metals haven't merged at all:



Therefore, the mass is simply the sum of the two: $1\cdot 1 + 2\cdot 1 = 3$. It is also the area of the two rectangles under the graph of the density function l, which is a step-function, and, therefore, the integral of l over $[0,2]$.

Exercise 3.2.2

What if the two rods have lengths 0.5 and 1.5?

Instead of just pointing out what m is, let's start from scratch.

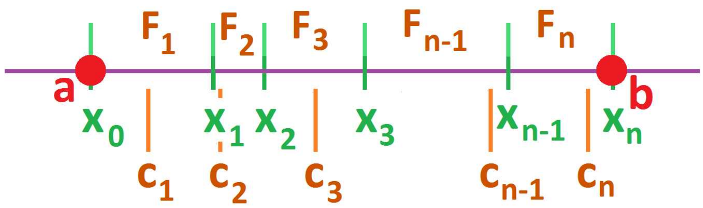
We have an augmented partition P:

$$a = x_0 \le c_1 \le x_1 \le \dots \le c_n \le x_n = b\,,$$

with these lengths of segments:

$$\Delta x_i = x_i - x_{i-1}\,.$$

We first imagine that the rod is divided into smaller pieces so that the density of each is found separately: $F_1, F_2, ..., F_n$:

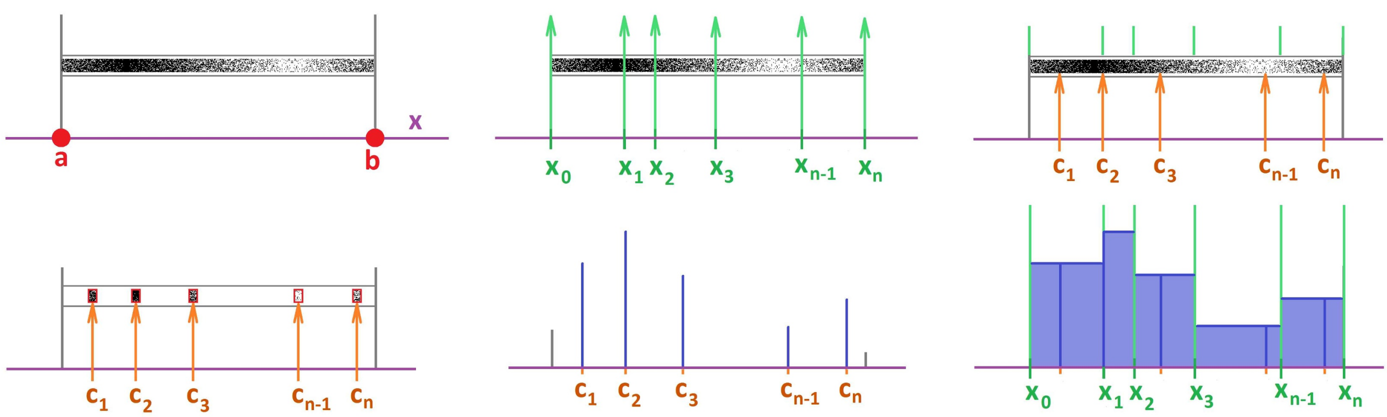


Then the total weight is simply the following:

$$\text{Total weight } = F_1\Delta x_1 + F_2\Delta x_2 + ... + F_n\Delta x_n\,.$$

The formula is sufficient for applications when an approximation is sufficient.

Here, we assume that the density is changing continuously and we cut the rod into these small segments by the planes starting at $x = x_i$ and then *sample* its density at the points c_i:



Then the density of each segment – if uniform – is given by $l(c_i)$ and we have:

$$\text{Mass of } i\text{th segment} = \text{density} \cdot \text{length} = l(c_i) \cdot \Delta x_i\,.$$

Then,

$$\text{Total mass } = \sum_{i=1}^{n} l(c_i) \cdot \Delta x_i\,.$$

We *recognize* this expression as the Riemann sum of the linear density function over this partition:

Definition 3.2.3: mass

If a function l defined at the secondary nodes of a partition of a segment $[a,b]$ is called *linear density*, then its Riemann sum

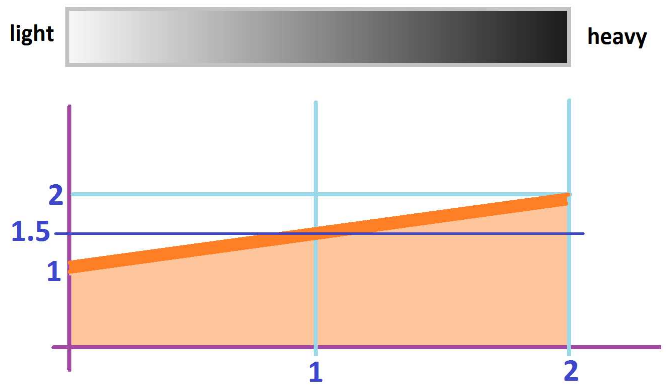
$$\Sigma\, l \cdot \Delta x$$

is called the *mass* of the segment.

Now, what if the density varies continuously?

Example 3.2.4: linear dependence

Suppose the density of a rod of length 2 is changing linearly: from 1 to 2. Then the meaning of the average density is clear; it is 1.5. We just average. Let's make sure that the definition makes sense:



It follows that the mass is $1.5 \cdot 2 = 3$. It is also the area of the trapezoid under the graph of the density function $l(x) = 1 + x/2$ and, therefore, the integral of this function over $[0, 2]$.

If the density is variable, then the mass of each segment – when short enough – is approximated by the mass of such a segment made entirely of material of density $l(c_i)$:

$$\text{Mass of } i\text{th segment} \approx \text{density} \cdot \text{length} = l(c_i) \cdot \Delta x_i \,,$$

and

$$\text{Total mass } \approx \sum_{i=1}^{n} l(c_i) \cdot \Delta x_i = \Sigma\, l \cdot \Delta x \,.$$

Then, we define the *mass* of the rod as the limit, if it exists, of these Riemann sums, i.e., the Riemann integral of l:

Definition 3.2.5: mass

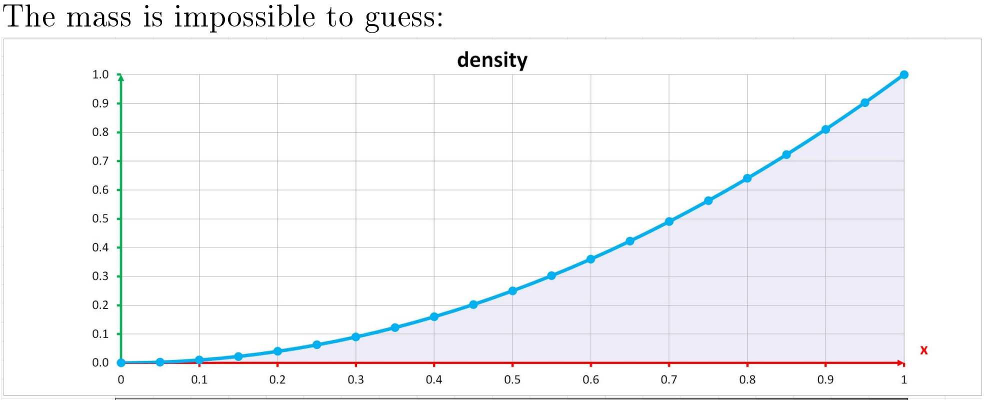
If an integrable function l on segment $[a, b]$ is called *linear density*, then its Riemann integral

$$\int_a^b l\, dx$$

is called the *mass* of the segment.

Example 3.2.6: quadratic dependence

Suppose we have again a rod of length 2 with the density changing from 1 to 2, but quadratically, $l(x) = x^2 + 1$. The mass is impossible to guess:

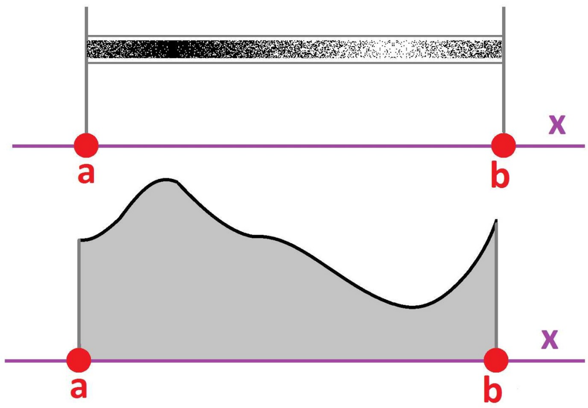


density, l(x)	0.00	0.00	0.01	0.02	0.04	0.06	0.09	0.12	0.16	0.20	0.25	0.30	0.36	0.42	0.49	0.56	0.64	0.72	0.81	0.90	1.00
location, x	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00

We compute the integral:

$$\text{Mass } = \int_a^b l\, dx = \int_0^1 (x^2 + 1)\, dx = \frac{x^3}{3} + x\bigg|_0^1 = \frac{4}{3} \,.$$

Here is another way to explain our definition. We realize that every location with higher density simply contains *more material* and we can just spread it out – vertically – making a *plate* that is wider at this spot and thinner at the location with a lower density:



In reverse, imagine that the area under the graph is made of a sheet of metal, which is then *rolled* into a non-uniform rod.

Exercise 3.2.7

Find how the mass of a rod with an exponentially growing density grows.

The *Fundamental Theorem of Calculus* provides further insight. Suppose $m(x)$ is the weight of the rod from a to x. Then the derivative of this function is the density:

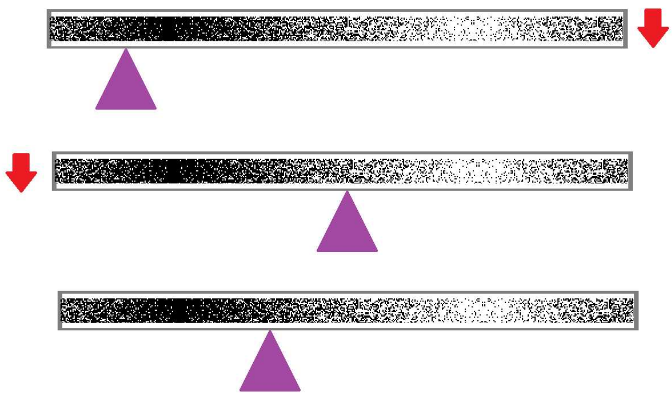
$$m'(x) = l(x)\,.$$

Exercise 3.2.8

Is it meaningful to speak of the mass of an infinitely long rod?

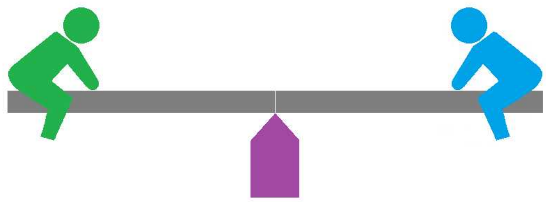
3.3. *The center of mass*

Can we now balance this non-uniform rod on a single point of support? Trial and error suggest this:

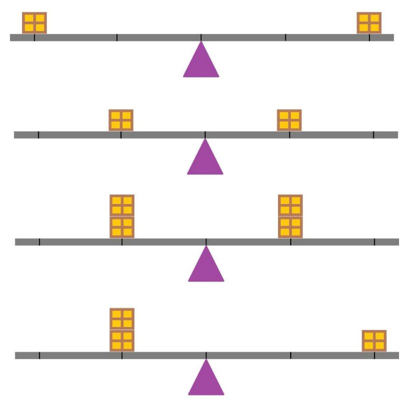


The question is important because this point, called the *center of mass*, is the center of rotation of the object.

The analysis starts with a simplest case, *seesaw*. Two persons of equal weight will be in a state of balance when located at equal distances from the point of support:



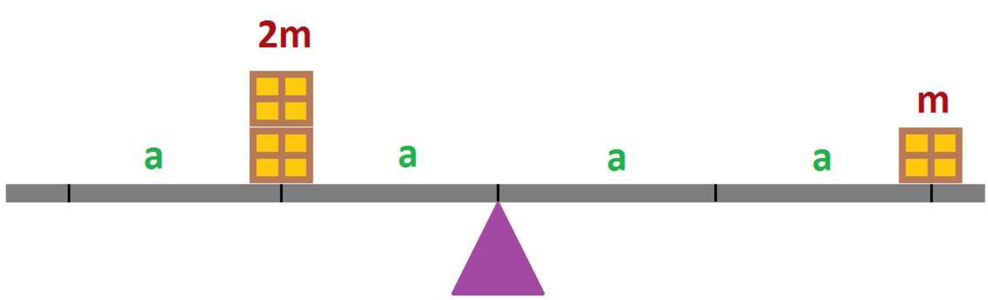
Now, what can be changed? What if one person is heavier than the other? From experience, we know that the former person should sit farther from the center in order to balance the beam:



In fact, if the person is twice as heavy as the other, the distance for the other should be twice as long! Conversely, if one person sits farther from the center than the other of the same weight, the former person should be joined by another in order to balance the beam.

Suppose the shorter distance is a and the smaller weight is m. Then, combined, the distances are a and $2a$ and weights are $2m$ and m. We express this data via the *balance equation*:

$$(a)(2m) = (2a)(m)\,.$$



In other words, this expression:

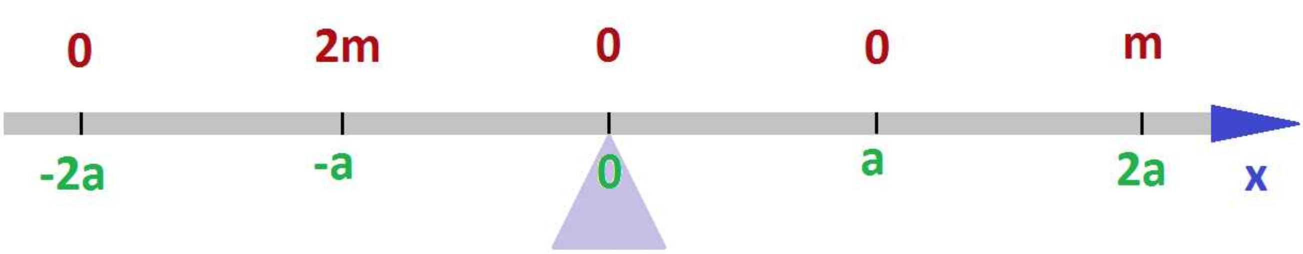
$$\text{distance} \ \cdot \ \text{weight}$$

called *the moment*, is the same to the left and to the right of the support. This distance is also called the *lever*.

Let's add the x-axis.

We then realize that it is the *signed distance*, i.e., the x-coordinate, of the object that matters. We simply re-write the balance equation:

$$(-a)(2m) + (2a)(m) = 0\,.$$



Then,

$$\text{moment} \;=\; \text{coordinate} \;\cdot\; \text{weight}$$

Furthermore, we can assume there is an object at every location but the rest of them have 0 mass. The balance equation becomes:

$$... + (-2a)(0) + (-a)(2m) + (0)(0) + (a)(0) + (2a)(m) + ... = 0\,.$$

This analysis brings us to the idea of combining *the weights and the distances* in a proportional manner in order to evaluate the contribution of a particular weight to the overall balance. The balance equation simply says that the sum of all moments is 0.

Definition 3.3.1: weights

We call a *system of weights* any collection of non-negative numbers $m_1, ..., m_n$ called *weights* assigned to n locations with coordinates $a_1, ..., a_n$ on the x-axis.

Definition 3.3.2: total moment

The *total moment of the system of weights with respect to the origin* is defined to be the sum of the moments of the weights, i.e.,

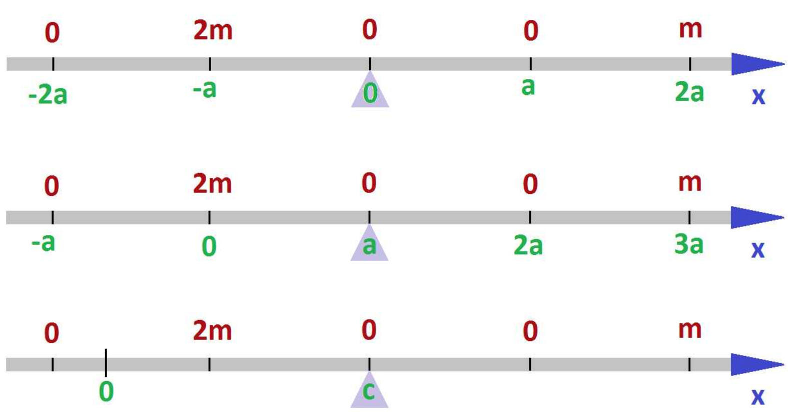
$$\sum_i m_i a_i\,.$$

The *balance equation* of the system states that its total moment is zero.

We now go back to the original problem:

► Suppose different weights are located on a beam, where do we put the support in order to balance it?

It was entirely our decision to place the origin of our x-axis at the center of mass. The result we have established should be independent from that choice and we can move the origin anywhere.

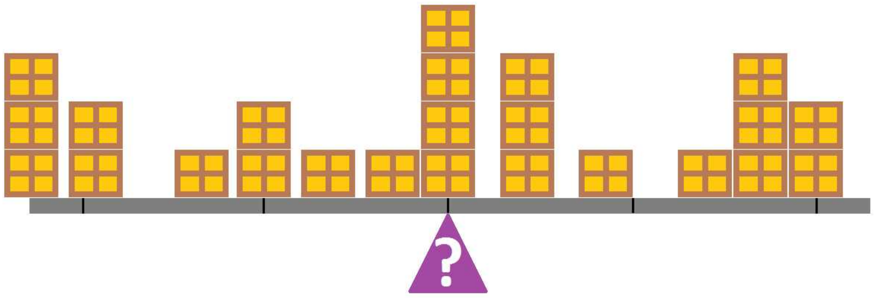


We just need to execute a *change of variables*. Suppose the center of mass (and the origin of the old coordinate system) is located at the point with coordinate c of the new coordinate system. Then, the new coordinate of the ith object is

$$c_i = a_i + c\,.$$

Therefore, the balance equation has this form:

$$\sum_i m_i(c_i - c) = 0\,.$$



We rewrite:

$$\sum_i m_i c_i = c \sum_i m_i \,.$$

The meaning of the equation may be seen as follows:

- ▶ The whole weight is concentrated at c.

Hence the name.

Definition 3.3.3: moment

Suppose we have a system of weights $m_1, ..., m_n$ located at $c_1, ..., c_n$ on the x-axis. For a given point c and for each i, the product

$$m_i(c_i - c)$$

is called the ith weight's *moment with respect to* c. The sum of the moments,

$$\sum_i m_i(c_i - c)\,,$$

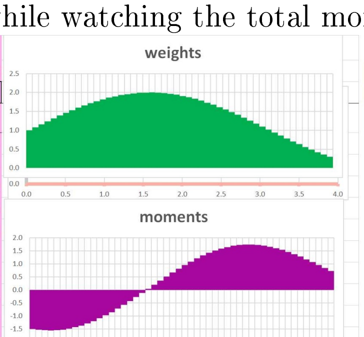
is called the *total moment with respect to* c. The *center of mass of this system of weights* is such a point c that the total moment with respect to c is zero.

Of course, if $c = 0$, we have the old definition.

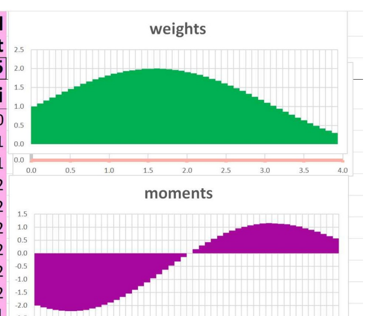
Example 3.3.4: center of mass by trial and error

Following this insight, let's find the center of mass of an object. The method amounts to trial and error. We just move c while watching the total moment:

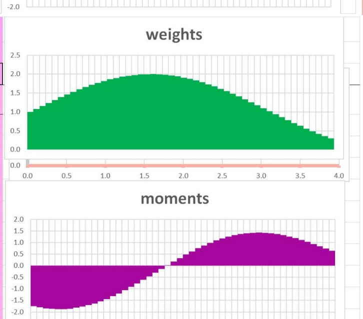
C	M		Total moment
1.5	71.5		17.3
Ci	**Mi**	**Li**	**MiLi**
0.0	1.0	-1.5	-1.5
0.1	1.1	-1.4	-1.5
0.2	1.2	-1.3	-1.6
0.2	1.2	-1.3	-1.6
0.3	1.3	-1.2	-1.6
0.4	1.4	-1.1	-1.5
0.5	1.5	-1.0	-1.5
0.6	1.5	-0.9	-1.4
0.6	1.6	-0.9	-1.4
0.7	1.7	-0.8	-1.3



C	M		Total moment
2.0	71.5		-18.5
Ci	**Mi**	**Li**	**MiLi**
0.0	1.0	-2.0	-2.0
0.1	1.1	-1.9	-2.1
0.2	1.2	-1.8	-2.1
0.2	1.2	-1.8	-2.2
0.3	1.3	-1.7	-2.2
0.4	1.4	-1.6	-2.2
0.5	1.5	-1.5	-2.2
0.6	1.5	-1.4	-2.2
0.6	1.6	-1.4	-2.2
0.7	1.7	-1.3	-2.1



C	Total mass		Total moment
1.75	71.5		-0.6
Ci	**Mi**	**Li**	**MiLi**
0.0	1.0	-1.8	-1.8
0.1	1.1	-1.7	-1.8
0.2	1.2	-1.6	-1.8
0.2	1.2	-1.5	-1.9
0.3	1.3	-1.4	-1.9
0.4	1.4	-1.4	-1.9
0.5	1.5	-1.3	-1.9
0.6	1.5	-1.2	-1.8
0.6	1.6	-1.1	-1.8
0.7	1.7	-1.0	-1.7



Exercise 3.3.5

What if we allow the values of m_i to be negative? What is the meaning of the system and of c?

To make our task easier, we solve the balance equation for c.

Theorem 3.3.6: Center of Mass

If c is the coordinate of the center of mass of the system of weights, then we have:

$$\boxed{c = \frac{\sum_i m_i c_i}{\sum_i m_i}}$$

Exercise 3.3.7

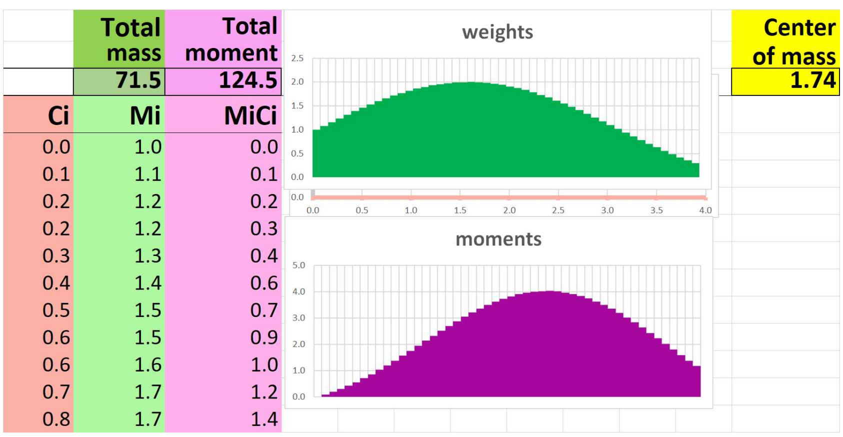
Prove the formula.

In other words, we have:

$$\boxed{\text{Center of mass } = \frac{\text{total moment}}{\text{total mass}}}$$

Example 3.3.8: center of mass from the formula

Armed with this formula, we can quickly find the centers of mass of objects. Below is the shape from the last example:



	Total mass	Total moment
	71.5	124.5
Ci	**Mi**	**MiCi**
0.0	1.0	0.0
0.1	1.1	0.1
0.2	1.2	0.2
0.2	1.2	0.3
0.3	1.3	0.4
0.4	1.4	0.6
0.5	1.5	0.7
0.6	1.5	0.9
0.6	1.6	1.0
0.7	1.7	1.2
0.8	1.7	1.4

Center of mass
1.74

Example 3.3.9: two objects

Let's test the formula for a system of just two objects.

First, suppose we have two identical weights located at a and b. Then

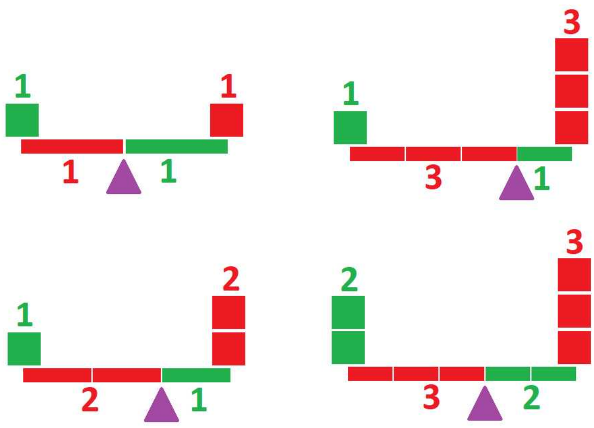
$$c = \frac{ma + mb}{m + m} = \frac{a + b}{2}\,.$$

So, no matter what the weight is, the center of mass lies halfway between the two objects, as expected.

What if the weights are different? We can guess that the center of mass will be closer to the heavier object. But by how much? Suppose these are m and $2m$. We compute:

$$c = \frac{ma + 2mb}{m + 2m} = \frac{a + 2b}{3} = \frac{1}{3}a + \frac{2}{3}b\,.$$

It's twice as close to the heavier object (bottom left):



In general, *the proportion of the distance is the proportion of the weight.*

Exercise 3.3.10

If α and β are the shares of the total weight, where is the center of mass of this two-object system?

More generally, the coordinate of the center of mass of the system can be re-written:

$$c = \frac{\sum_i m_i c_i}{\sum_i m_i} = \sum_i \frac{m_i}{\sum_j m_j} c_i \, .$$

Therefore, we have the following.

Corollary 3.3.11: Weighted Average

If c is the coordinate of the center of mass of a system of weights with locations at c_i, weights m_i, and the total weight M, then

$$c = \sum_i \mu_i c_i \, ,$$

where μ_i are the relative weights:

$$\mu_i = \frac{m_i}{M} \, .$$

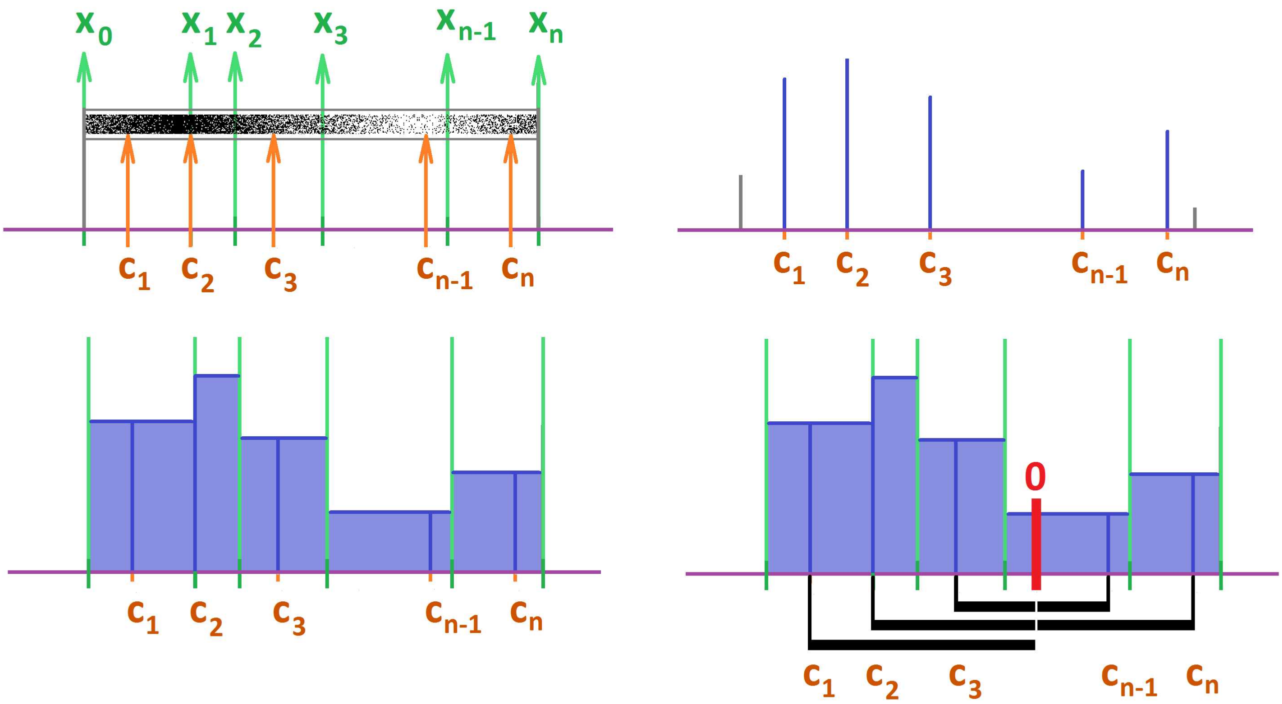
We start to notice that the numerous blocks placed on the bar start to look like the graph of a function! The value of this function is the height of the blocks placed at that location. We know that this function can also be seen as the linear density of a rod.

Next, let's imagine that the density varies in a more unpredictable way.

We continue in the same way as in the last section – an augmented partition P of interval $[a, b]$ is given:

$$a = x_0 \le c_1 \le x_1 \le ... \le x_{n-1} \le c_n \le x_n = b$$

Then the density function l is defined at the secondary nodes. Then the terms $l(c_i)\Delta x_i$ representing the weight of each interval are formed... but not simply added this time:



Each of these terms is a weight placed on top of the interval visualized as a rectangle. However, it is assumed that the weight of the ith rectangle is concentrated at c_i. The lever of each weight is also shown. Then the total moment of this system of weights with respect to some c is the following:

$$\sum_i m_i(c_i - c) = \sum_i l(c_i) \cdot \Delta x_i(c_i - c) = \sum_i l(c_i)(c_i - c) \cdot \Delta x_i \,.$$

Have we produced a Riemann sum as before? Well, this isn't the Riemann sum of l! Let's try this function (dependent on our choice of c):

$$f(x) = l(x)(x - c) \,.$$

Then, indeed, we face its Riemann sum:

$$\sum_i m_i(c_i - c) = \Sigma\, f \cdot \Delta x \,.$$

Just as above, the system of weights that makes up the rod is balanced when the total moment is zero:

$$\Sigma\, l(x)(x - c) \cdot \Delta x = 0 \,.$$

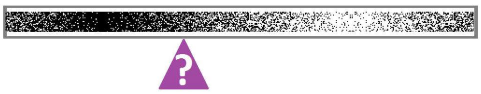
We arrive to a similar conclusion below.

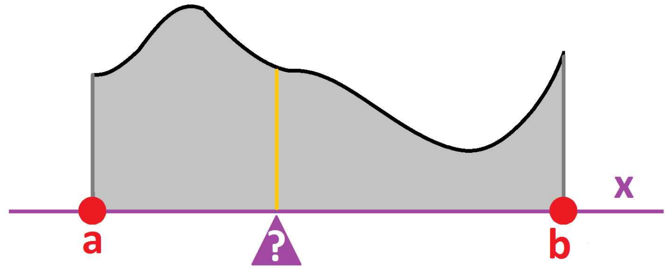
Theorem 3.3.12: Center of Mass – Discrete Case

Suppose a function $y = l(x)$ is defined at the secondary nodes c_i, $i = 1, 2, ..., n$, of a partition of interval $[a, b]$. Then the system of weights $l(c_i)\Delta x_i$, $i = 1, 2, ..., n$, has its center of mass at the following point:

$$\boxed{c = \frac{\Sigma\, l(x)x \cdot \Delta x}{\Sigma\, l(x) \cdot \Delta x}}$$

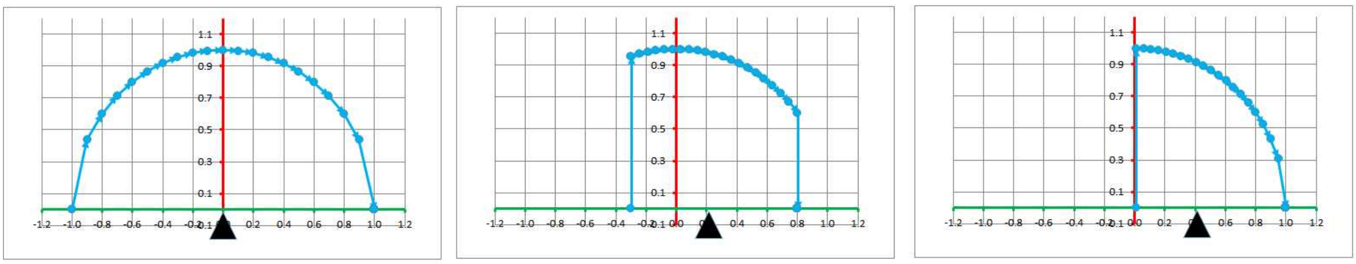
What we have discovered is that the problem of balancing a rod with a variable density is equivalent to the problem of balancing the region below the graph of the density function:





Example 3.3.13: piece of circle

Let's test this formula on some regions cut from the unit circle:



In Chapter 4, we will offer more complex examples.

Exercise 3.3.14

Prove that if the density of the rod is strictly increasing (or decreasing), its center of mass cannot be in the center.

The next step is to think of the weights assigned to *every* location on the x-axis. In other words, the distribution of weight is no longer incremental.

What we have learned is that the total moment of the region with respect to some c is approximated by that of this system of weights, which is the Riemann sum,

$$\sum_i m_i(c_i - c) = \sum_i l(c_i) \cdot \Delta x_i (c_i - c) = \Sigma\, f \cdot \Delta x\,,$$

of the function

$$f(x) = l(x)(x - c)\,.$$

The beam doesn't have to be balanced and the total moment doesn't have to be zero *for each partition*, but it does have to diminish to zero as we refine the partitions. This means that the Riemann integral of this function is zero.

Definition 3.3.15: center of mass

Suppose we have a non-negative function $y = l(x)$ integrable on segment $[a, b]$ called the *density function*. Then, for a given point c, the integral

$$\int_a^b l(x)(x - c)\, dx$$

is called the *total moment of the segment with respect to* c. The *center of mass of the segment with density function* l is such a point c that the total moment with respect to c is zero.

Just as in the discrete case, the balance equation can be solved for c:

Theorem 3.3.16: Center of Mass – Continuous Case

Suppose we have a non-negative function $y = l(x)$ integrable on interval $[a, b]$. If the mass of the segment is not zero, then the center of mass is:

$$\boxed{c = \frac{\int_a^b l(x)x\,dx}{\int_a^b l(x)\,dx}}$$

Proof.

First, we note that $y = l(x)(x - c)$ is integrable by PR. Then we use *SR* and *CMR* to compute the following:

$$0 = \text{ total moment } = \int_a^b l(x)(x-c)\,dx = \int_a^b l(x)x\,dx + c\int_a^b l(x)\,dx\,.$$

Now solve for c.

Once again, we have:

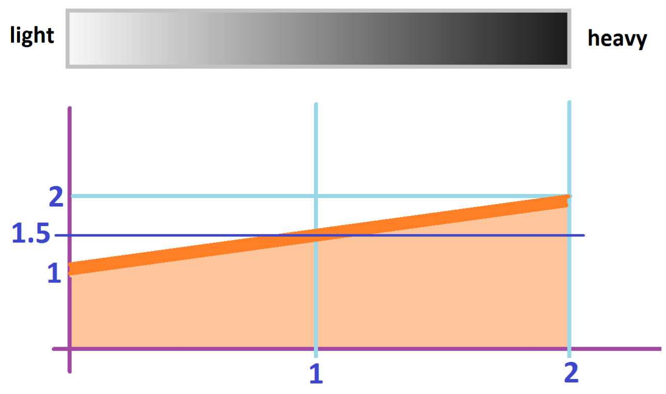
$$\boxed{\text{Center of mass } = \frac{\text{total moment}}{\text{total mass}}}$$

Exercise 3.3.17

Show this theorem implies the previous one.

Example 3.3.18: linear dependence

Suppose the density of a rod of length 2 is changing linearly: from 1 to 2, i.e., $l(x) = x/2 + 1$.



Then, the mass is 3. It was found in the last section based on a common sense analysis. That's the denominator of the fraction. Now, the numerator. Mere common sense won't help this time; we need to integrate:

$$\begin{aligned}
\int_0^2 l(x)x\,dx &= \int_0^2 (x/2+1)x\,dx \\
&= \int_0^2 (x^2/2+x)\,dx \\
&= x^3/6 + x^2/2\Big|_0^2 \\
&= 8/6 + 4/2 \\
&= 10/3\,.
\end{aligned}$$

Therefore, the center of mass is

$$c = \frac{10}{3} \div 3 = \frac{10}{9}\,.$$

Slightly to the right of the center...

Exercise 3.3.19

Find the center of mass of a rod with a linearly increasing density.

Exercise 3.3.20

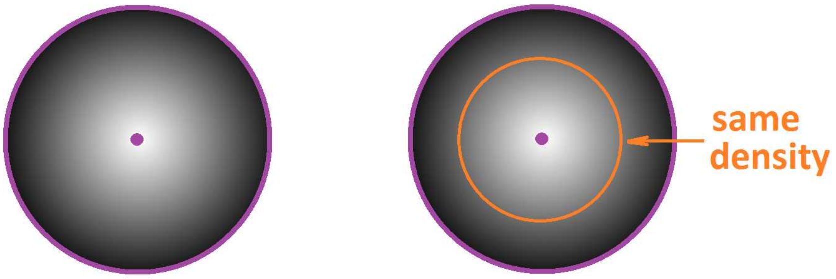
Find the center of mass of a plate cut from a circle of radius 1 centered at the origin by the lines $x = a$ and $x = b$.

Exercise 3.3.21

What is the meaning of the center of mass of an infinite object?

3.4. The radial density and the mass

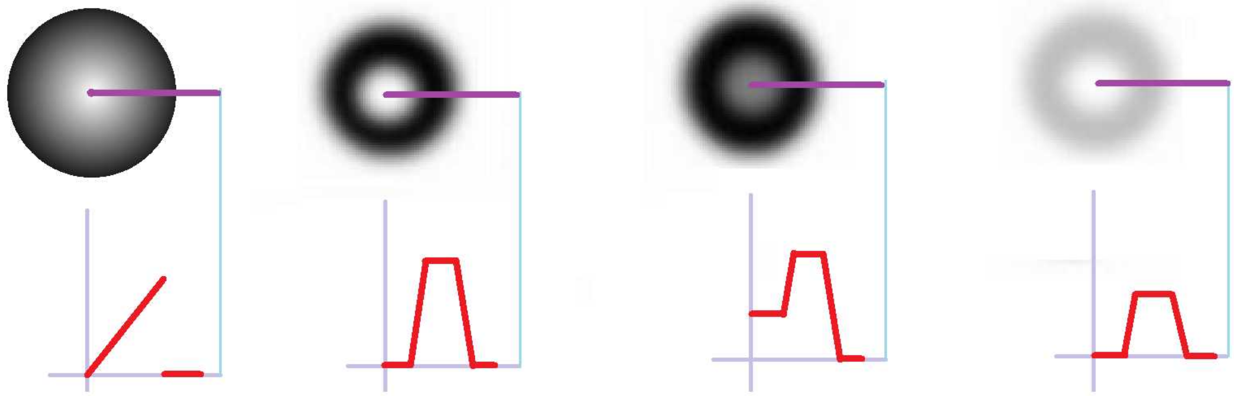
Suppose next we have an alloy that is rotated as it hardens. Then its density depends (only) on the distance from the center.



The same effect is produced by stirring a liquid.

In either case, we ignore the depth and all we see is a disk. Then, for any radial line (we pick one and call it the x-axis) there is no change in density in the directions perpendicular to it. We then ignore those directions and the density becomes a function of a single number x designating the distance to the center along this line; hence the *radial* density $y = r(x)$.

Here are a few examples of this function:

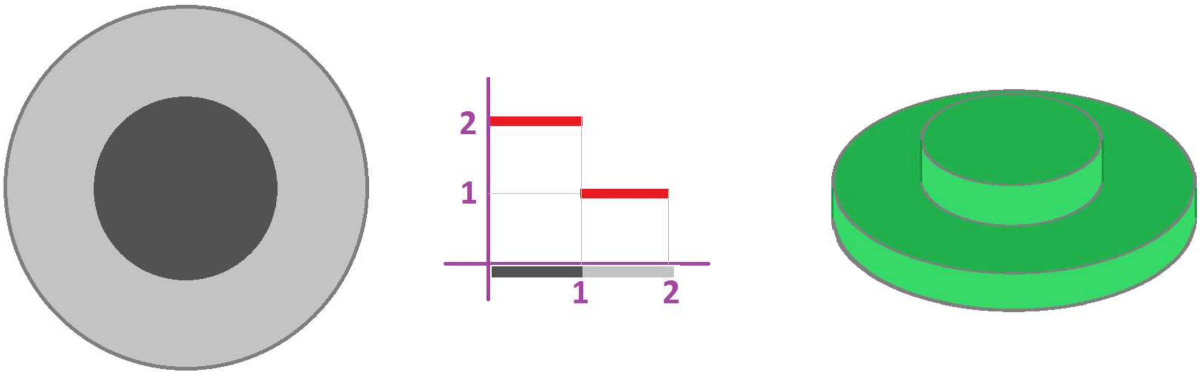


We will provide analysis similar to the one in the last section to define the mass of such an object.

Suppose the radial density r is given, what is the mass of the disk?

Example 3.4.1: two pieces

Suppose the two metals with densities 2 on the inside and 1 on the outside haven't merged at all. The object is simply a combination of a disk of radius 1 and a washer around it of thickness 1:



Then, the mass is simply the sum of the mass of the disk and the mass of the washer:

$$\begin{aligned}\text{Mass} &= 2\cdot\text{area of the disk} && +1\cdot\text{area of the washer}\\ &= 2\cdot\pi\cdot 1^2 && +1\cdot\left(\pi\cdot 2^2-\pi\cdot 1^2\right).\end{aligned}$$

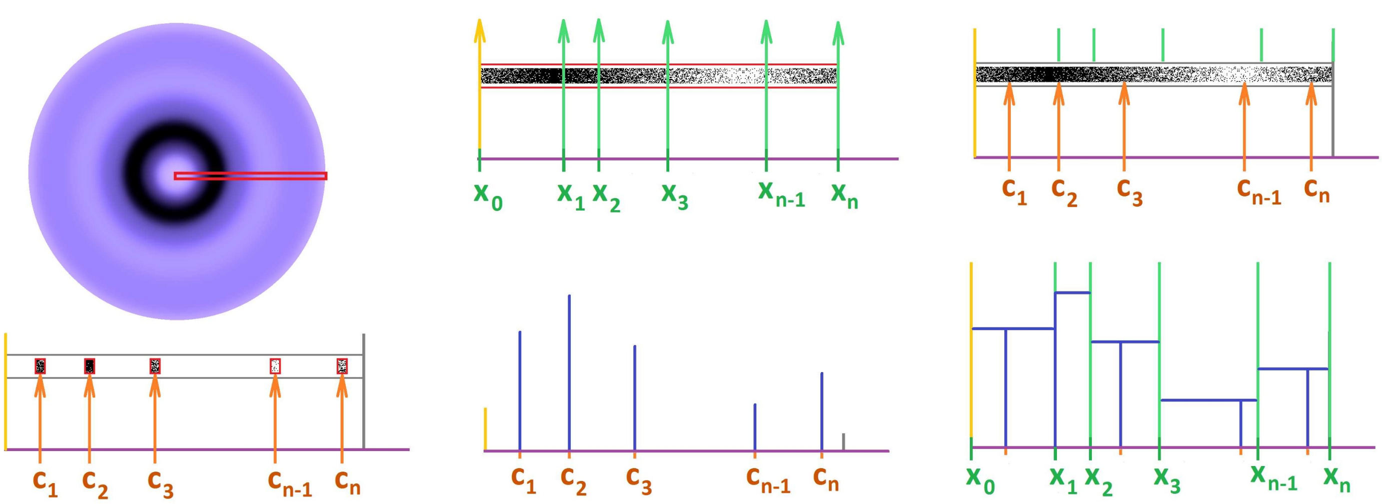
It's 5π.

We can just replace the disk that has a constant thickness and a variable density with one that has a variable thickness and a constant density. Then we can use the results of the last section. Instead we start from scratch.

Suppose we have an augmented partition P of the radius:

$$a=x_0\le c_1\le x_1\le ...\le c_n\le x_n=b$$

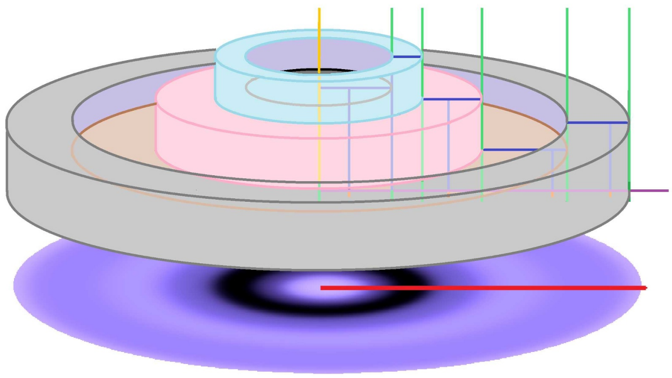
Here, we cut the disk into small washers by the cylinders starting at $x=x_i$ and then sample its density at the points c_i:



Then the density of each washer – when uniform – is $r(c_i)$ and we have:

$$\text{Mass of } i\text{th washer } = \text{density}\cdot\text{area} = r(c_i)\cdot\left(\pi x_i^2-\pi x_{i-1}^2\right)\,,$$

since the inside radius of the washer is x_{i-1} and the outside is x_i.



Then we have:

$$\boxed{\text{Mass of the disk } = \sum_{i=1}^{n} r(c_i) \cdot \pi \left(x_i^2 - x_{i-1}^2 \right)}$$

This formula is fine for computations but it is *not* the Riemann sum of any function!

A clever trick is to choose the secondary nodes to be the mid-points:

$$c_i = \frac{1}{2}(x_i + x_{i-1}) \, .$$

Then, we can factor the difference of two squares and simplify:

$$\text{Mass of the disk } = \sum_{i=1}^{n} r(c_i) \cdot \pi (x_i + x_{i-1})(x_i - x_{i-1}) = 2\pi \sum_{i=1}^{n} r(c_i) c_i \cdot \Delta x_i \, .$$

This *is* the Riemann sum of a simple function:

$$\text{Mass of the disk } = 2\pi \Sigma \, x r(x) \cdot \Delta x \, .$$

With this discovery, we can address the question: What if the *density varies continuously*?

Then the mass of each washer – when thin enough – is *approximated* by the mass of such a washer made entirely of material of density $r(c_i)$:

$$\text{Mass of } i\text{th washer } \approx \text{density} \cdot \text{area} = r(c_i) \cdot \left(\pi x_i^2 - \pi x_{i-1}^2 \right) \, .$$

Then we go through the same algebra:

$$\text{Total mass } \approx \sum_{i=1}^{n} r(c_i) \cdot \pi \left(x_i^2 - x_{i-1}^2 \right) = 2\pi \sum_{i=1}^{n} r(c_i) c_i \cdot \Delta x_i \, .$$

Then, we define the *mass* of the disk as the limit of these Riemann sums; i.e., the Riemann integral:

Definition 3.4.2: mass of disk

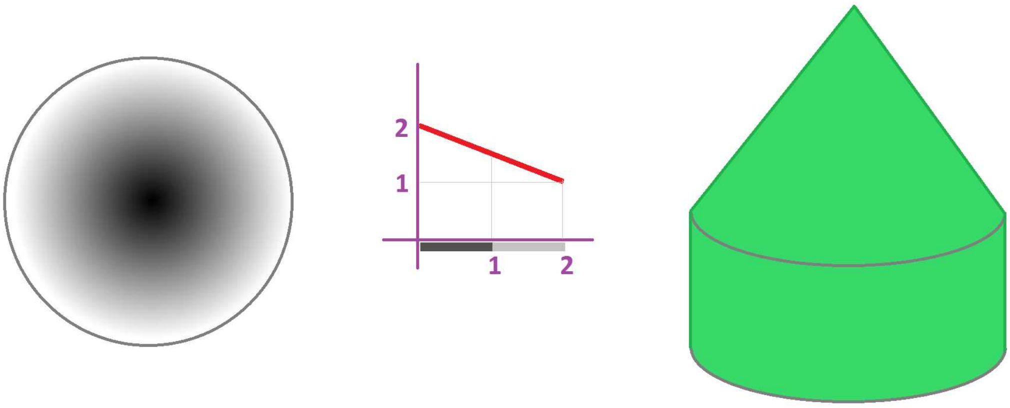
If an integrable function r on segment $[0, b]$ is called a *radial density*, then the following integral is called the *mass* of the disk of radius b:

$$\text{Mass } = 2\pi \int_0^b x r(x) \, dx \, .$$

Once again, we realize that each location with higher density simply contains more material and we can just spread it out – vertically – making the disk thicker at this spot and thinner at the location of lower density.

Example 3.4.3: linear dependence

Suppose the density of a disk of radius 2 is changing linearly: from 1 to 2. Then the meaning of the average density depends on the respective areas, as shown above.

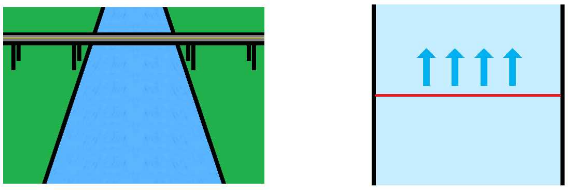


The mass must have something to do with the volume of this surface of revolution... Let's integrate:

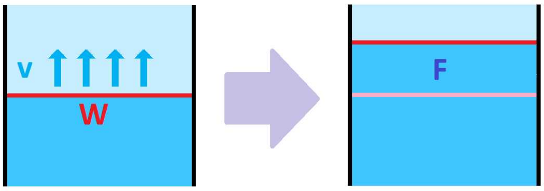
$$\begin{aligned}\text{Mass} &= 2\pi \int_a^b x r(x)\, dx \\ &= \pi \int_0^2 x(2 - x/2)\, dx \\ &= \pi \int_0^2 (2x - x^2/2)\, dx \\ &= \pi(x^2 - x^3/6)\Big|_0^2 \\ &= \pi(2^2 - 2^3/6) \\ &= \frac{8\pi}{3}\,.\end{aligned}$$

3.5. *The flow velocity and the flux*

Suppose water flows in a canal:



How much water is crossing the given line per unit of time? We will ignore the depth and consider this view from above:

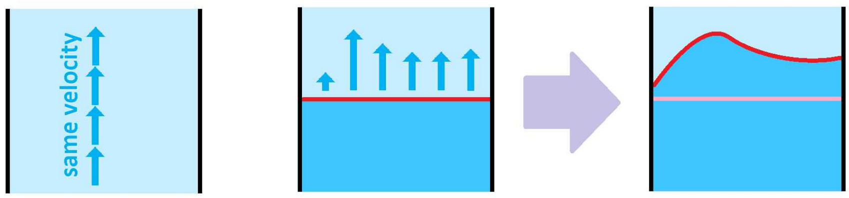


When the velocity of the water is the same at all locations, the total amount of the water that has crossed

the line, called the *flux* F, is the velocity v times the width W of the cross-section:

$$F = v \cdot W \,.$$

The flow velocity may vary depending on the location (not time!). We assume that the velocity is the same along the lines parallel to the walls of the canal. We visualize the process by imagining that a narrow strip of red dye is applied across the canal and then after, say, one minute we see how the die has progressed:



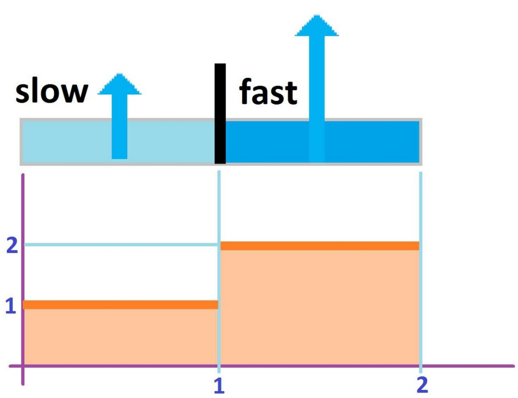
What is the flux then?

To begin with, we assume that the flow velocity depends on a *single* variable (one degree of freedom again), the location distance across the canal. Then, there is a line – we choose it to be interval $[a, b]$ on the x-axis – with no change in velocity in the directions perpendicular to it. Then the velocity is a function $y = v(x)$ of a single number x in $[a, b]$.

Suppose this time the velocity v is given as a function of location, what is the flux?

Example 3.5.1: two gates

Suppose we have two separate canals side by side, with the velocities 1 and 2 and the same width 1:



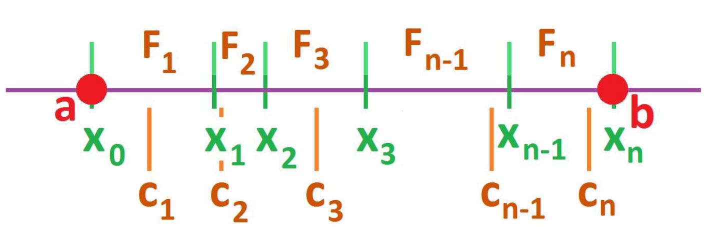
Therefore, the volume is simply the sum of the two: $1 \cdot 1 + 2 \cdot 1 = 3$. It is also the area of the two rectangles under the graph of the velocity function v, which is a step-function, and, therefore, the integral of l over $[0, 2]$.

Instead of just pointing out that the flux is an antiderivative of the speed (with respect to location not time!), let's start from scratch.

We have an augmented partition P:

$$a = x_0 \leq c_1 \leq x_1 \leq ... \leq c_n \leq x_n = b$$

We first imagine that the canal is divided into channels or locks so that the flow velocity through each is found separately: $F_1, F_2, ..., F_n$:



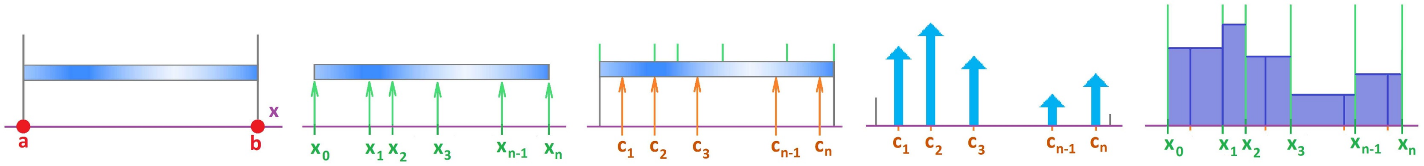
Then the *total flow* is simply

$$\text{total volume } = F_1\Delta x_1 + F_2\Delta x_2 + ... + F_n\Delta x_n\,.$$

The formula is sufficient for practical applications.

To find a formula for an idealized situation, we continue on.

We would like to find a Riemann sum here. We imagine that the flow velocity varies incrementally over the gates that divide the canal's cross-section. The canal is cut into segments by the line starting at $x = x_i$ and sampled velocity at the points c_i is $v(c_i)$:



Then we have:

$$\text{Flux though } i\text{th segment} = \text{velocity} \cdot \text{width} = v(c_i) \cdot \Delta x_i\,.$$

Then,

$$\text{Total flux } = \sum_{i=1}^{n} v(c_i) \cdot \Delta x_i\,.$$

We *recognize* this expression as the Riemann sum, $\Sigma\, v \cdot \Delta x$, of the velocity function over this partition.

Definition 3.5.2: flow velocity

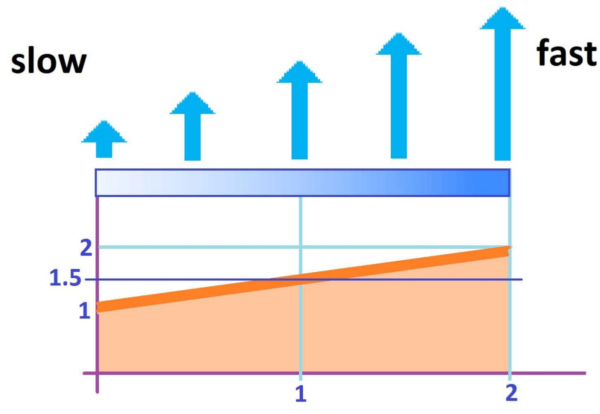
If a function v defined at the secondary nodes of a partition of a segment $[a, b]$ is called a *flow velocity*, then its Riemann sum is called the *flux*:

$$\text{Flux } = \Sigma\, v \cdot \Delta x\,.$$

What if the flow velocity varies continuously?

Example 3.5.3: linear dependence

Suppose the velocity of the canal of width 2 is changing linearly: from 1 to 2. Then the meaning of the average velocity is clear; it is 1.5.



Therefore, the volume is $1.5 \cdot 1 = 1.5$. It is also the area of the triangle under the graph of the velocity function $v(x) = 1 + x$ and, therefore, the integral of v over $[0, 2]$.

Then the flux through each segment – when short enough – is *approximated* by the volume with the water moving entirely at the velocity $v(c_i)$:

$$\text{Volume of } i\text{th segment} \approx \text{velocity} \cdot \text{width} = v(c_i) \cdot \Delta x_i\,.$$

Then,

$$\text{Flux } = \text{Total volume } \approx \sum_{i=1}^{n} v(c_i)\cdot \Delta x_i\,.$$

We define the *flux* of the rod as the limit, if it exists, of these Riemann sums, i.e., the Riemann integral of v.

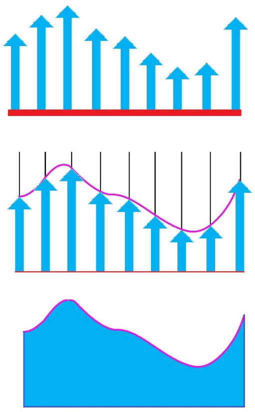
Definition 3.5.4: flux

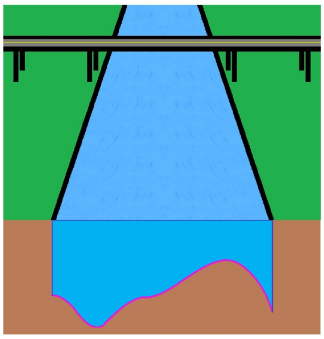
If an integrable function v on segment $[a,b]$ is called a *flow velocity* in a canal, then its Riemann integral is called the *flux* of the flow:

$$\text{Flux } = \int_a^b v\,dx\,.$$

Here is another way to explain this result. We can take our canal, with a variable water *velocity*, and imagine a canal with the same flux but a *constant* velocity. How is it possible? We think of each location with higher velocity as one that has *more water*.

The first approach is to spread the water out – vertically – making the canal *deeper* at this spot and shallower at the location with a lower velocity:



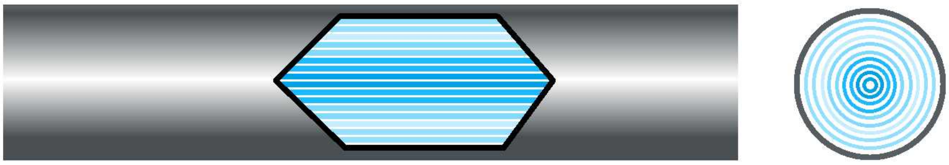


The second approach is to think of each location with higher velocity as simply one with *denser* liquid.

Exercise 3.5.5

What if this is an ocean, i.e., the cross-section of our "canal" is infinitely wide?

A variation of this analysis is as follows. Suppose now that the water flows through a *pipe*:



Suppose the flow velocity varies depending on the *distance* from the location to the pipe's wall. For example, the water may go slower next to the wall because of the friction. We have a circular pattern again...

Definition 3.5.6: flux

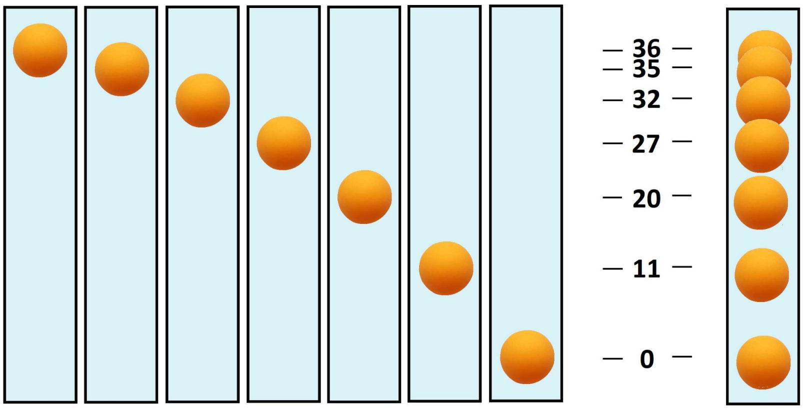
If an integrable function v on segment $[0,R]$ is called a *flow velocity* through a pipe of radius R, then the Riemann integral $2\pi\int_0^R xv(x)\,dx$ is called the *flux*.

Exercise 3.5.7

Following the ideas developed in this chapter, justify the above definition.

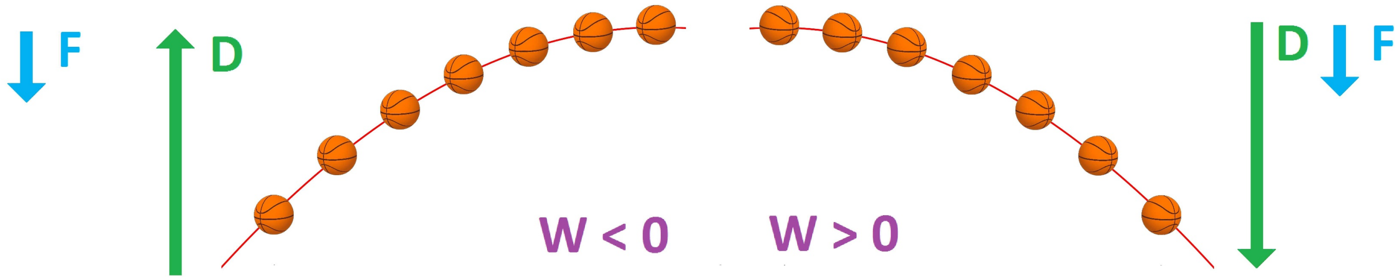
3.6. The force and the work

Suppose a ball is dropped on the ground from a certain height:



This phenomenon is the result of the gravitational force. This force is directed down, just as the movement of the ball. The work done on the ball by this force as it falls is equal to the (signed) magnitude of the force, i.e., the weight of the ball, multiplied by the (signed) distance to the ground, i.e., the displacement. All horizontal motion is ignored as unrelated to the gravity.

The need for using the *signed* distance D and force F is revealed by the example of moving an object up from the ground. Then the work W performed by the gravitational force is negative!



Of course, the sign in either case is determined by the direction of the axis we assign to the line of motion.

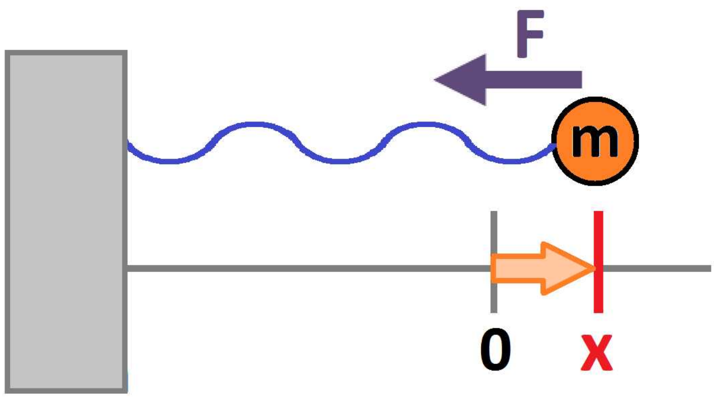
Suppose we are to move from point a on the x-axis to point $b > a$. When the force F is constant, the *work* W is equal to the force F times the distance covered between a and b:

$$W = F \cdot (b - a) .$$

The force may vary depending on the location between a and b.

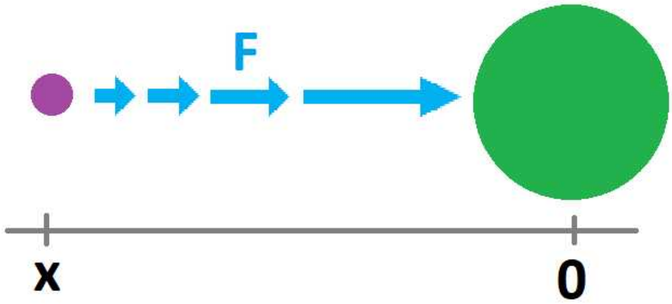
Example 3.6.1: physics

The examples of variable forces may be these: spring, gravitation, air pressure.



In the case of an object attached to a spring, the force is proportional to the (signed) distance of the object to its equilibrium according to the *Hooke's Law*:

$$F(x) = -kx\,.$$



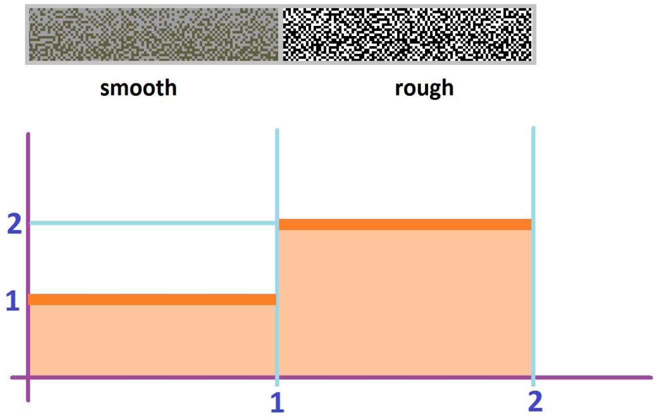
Away from the ground, the gravity is proportional to the reciprocal of the square of the distance of the object to the center of the planet according to the *Newton's Law of Gravitation*:

$$F(x) = -\frac{k}{x^2}\,.$$

The *pressure* and, therefore, the medium's resistance to motion may change arbitrarily.

Example 3.6.2: traction

Suppose the force is *traction*. Suppose that there are two distinct strips: one is smoother and the other rougher.



The force takes – between $a = 0$ and $b = 2$ – only two different values 1 and 2 switching at $c = 1$. Therefore, the work is simply the sum of the two over either of the segments: $1 \cdot 1 + 2 \cdot 1 = 3$. It is also the area of the two rectangles under the graph of the force function F, which is a step-function, and, therefore, the integral of l over $[0, 2]$.

For the general case, we have an augmented partition P:

$$a = x_0 \le c_1 \le x_1 \le ... \le c_n \le x_n = b$$

The path is divided into small segments by $x = x_i$ and then the force is sampled at the points c_i. Then the force on each segment – if constant – is equal to $F(c_i)$ and we have:

$$\text{Work on } i\text{th segment} = \text{ force } \cdot \text{ length} = F(c_i) \cdot \Delta x_i \,.$$

Then,

$$\text{Total work } = \sum_{i=1}^{n} F(c_i) \cdot \Delta x_i \,.$$

Once again, we *recognize* this expression as the Riemann sum, $\Sigma\, F \cdot \Delta x$, of the force function over this partition.

Definition 3.6.3: work

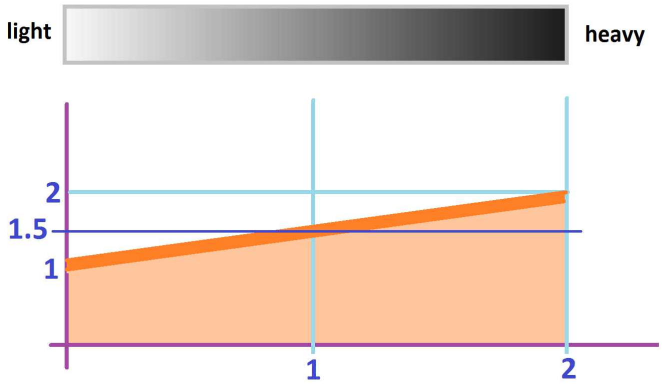
If a function F is defined at the secondary nodes of a partition of a segment $[a, b]$ is called a *force function*, then its Riemann sum is called the *work* of the force over interval $[a, b]$:

$$\text{Work } = \Sigma\, F \cdot \Delta x \,.$$

What if the force varies "continuously"?

Example 3.6.4: linear dependence

Suppose the force is changing linearly over the interval $[0, 2]$: from 1 to 2. Then the meaning of the average force is clear; it is 1.5.



Therefore, the work is $1.5 \cdot 2 = 3$. It is also the area of the triangle under the graph of the force function $F(x) = 1 + x/2$ and, therefore, the integral of this function over $[0, 2]$. When the change of the force is non-linear, the argument fails.

The work on each segment is approximated by the work with the force being constantly equal to $F(c_i)$:

$$\text{Work on } i\text{th segment} \approx \text{ force } \cdot \text{ length} = F(c_i) \cdot \Delta x_i \,.$$

Then,

$$\text{Total work } \approx \sum_{i=1}^{n} F(c_i) \cdot \Delta x_i \,.$$

We define the work of the force as the limit, if it exists, of these Riemann sums, i.e., the Riemann integral of F.

Definition 3.6.5: work

If an integrable function F on segment $[a, b]$ is called a *force function*, then its Riemann integral is called the *work* of the force over interval $[a, b]$:

$$\text{Work } = \int_a^b F \, dx \,.$$

Exercise 3.6.6

How much work does it take to move an object attached to a spring s units from the equilibrium?

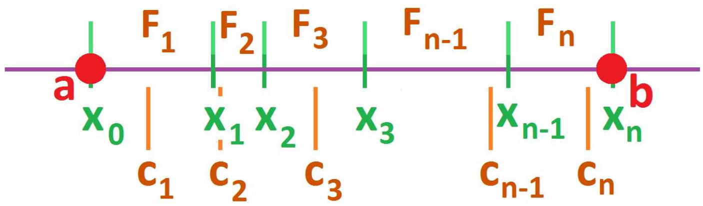
Exercise 3.6.7

How much work does it take to move an object s units from the center of a planet?

As a *summary*, we have solved the problem of finding a certain quantity W – work, flow, and mass – in an identical manner. We have an augmented partition P:

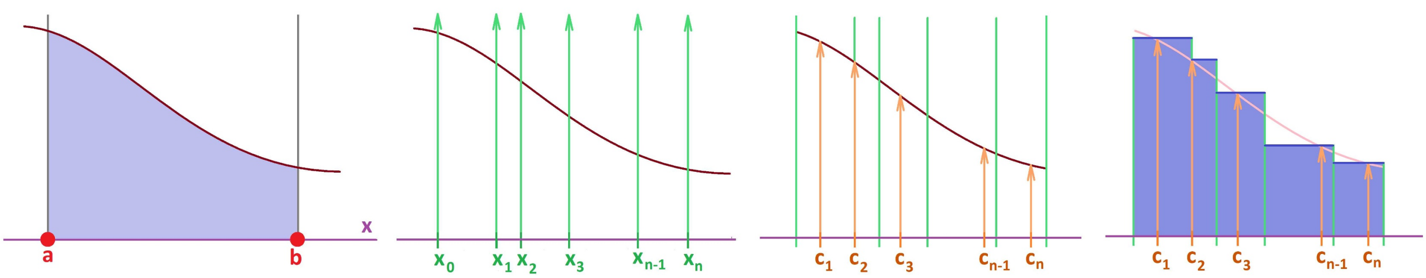
$$a = x_0 \le c_1 \le x_1 \le ... \le c_n \le x_n = b$$

We divide the path into small segments by $x = x_i$ and then sample quantity F – the force, or the flow speed, or the linear density – at the points c_i:



Then this quantity, $F(c_i)$, on each segment is used to find the value of W:

$$W_i = F(c_i) \cdot \Delta x_i \,.$$



Then, the total approximated value of W over the whole segment is

$$\sum_i W_i = \sum_{i=1}^{n} F(c_i) \cdot \Delta x_i \,,$$

which is the Riemann sum of F over this partition. The exact value of the total of W is the limit of these Riemann sums, i.e., the Riemann interval of F:

$$W = \int_a^b F\,dx \,.$$

Warning!

Contrary to the unified approach presented here, the treatments of the three integrals are substantially different in dimension 3 (Chapter 4HD-5 and 4HD-6):

- The work is an integral over a curve.
- The flow is an integral over a surface.
- The mass is an integral over a solid.

We now consider a different setup...

We arrived at the integral formula above because of a simple ("additive") property of work:

> ▶ When there are two segments of the trip, the work to move through the two is equal to the work required to move through the first plus the work required to move through the second.

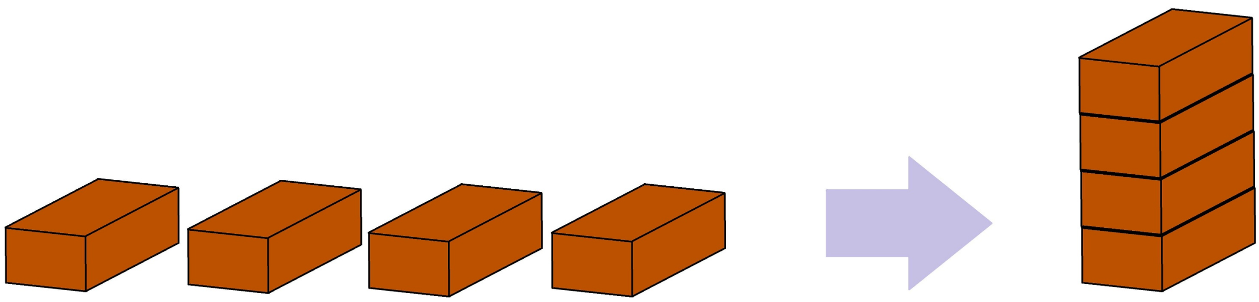
We are to consider a situation when

> ▶ Two objects, possibly identical, under a force, possibly constant, have to be moved *different distances*.

Then there is no such a shortcut formula.

Example 3.6.8: bricks

An example of such a task is stacking bricks:



Then the work – of the person acting *against* the gravity – is the following:

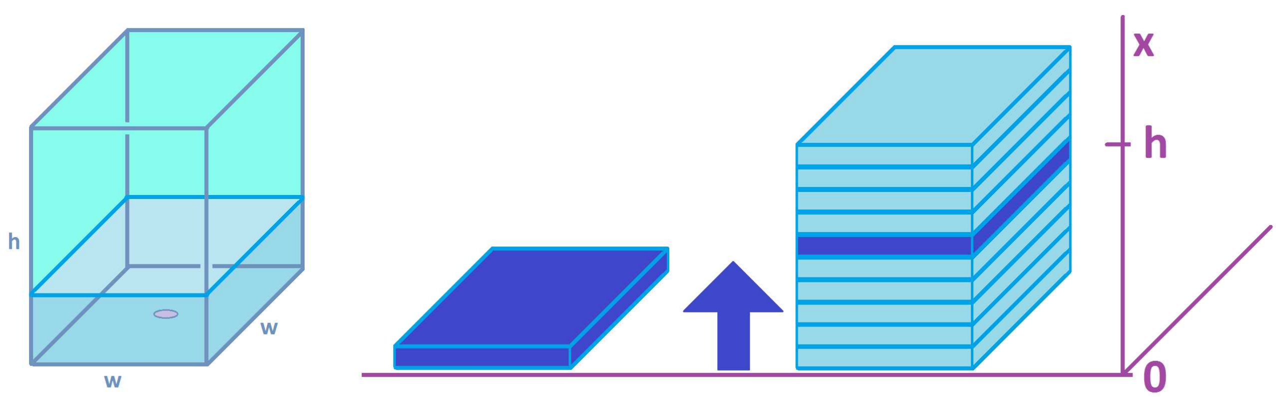
$$W = M \cdot 0 \cdot h + M \cdot 1 \cdot h + M \cdot 2 \cdot h + M \cdot 3 \cdot h \,,$$

where M is the weight of the brick and h is its height.

Furthermore, we consider the possibility that the force is constant but the object can't be thought of as a point anymore. In other words, different parts of the object will travel different distances. This situation isn't covered by the above definition of work.

Example 3.6.9: cubical tank

Suppose we are to fill a tank with $w \times w$ base and height h with water – from the bottom:



What is the work required assuming that the density is 1?

We imagine that water appears at the bottom in thin slices and then each is delivered to the appropriate height. They come from an augmented partition P of $[0, h]$. This means that the x-axis is vertical. The ith slice is a square between the planes $x = x_{i-1}$ and $x = x_i$. Its thickness is $\Delta x_i = x_i - x_{i-1}$ and its weight is $w^2 \cdot \Delta x_i$. Now, the ith slice is delivered to height c_i. The work to do so is

$$w^2 \Delta x_i \cdot c_i \,.$$

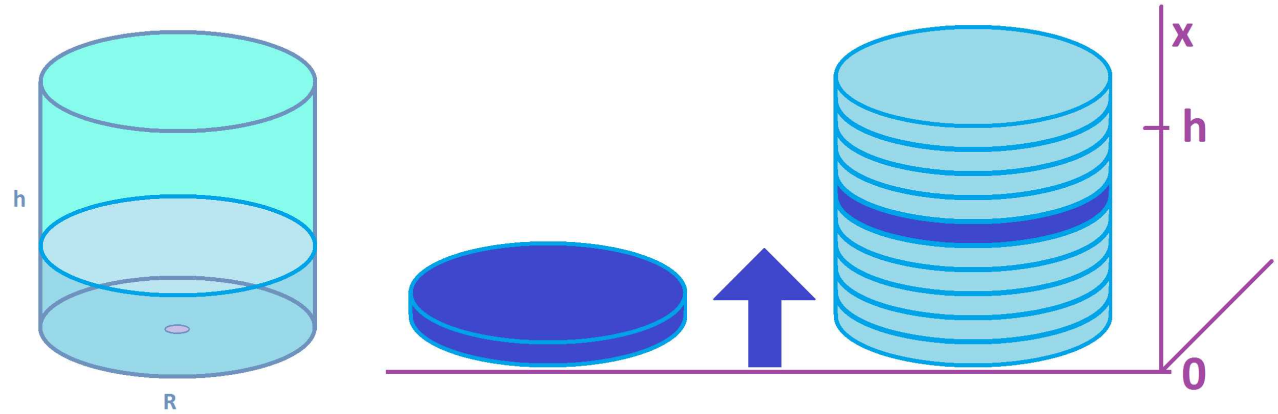
Then the total work is approximated by:

$$\text{Work } \approx \sum_{i=1}^{n} w^2 c_i \cdot \Delta x_i \,.$$

This is the Riemann sum of the integral:

$$\text{Work } = w^2 \int_0^h x\, dx = w^2 \frac{h^2}{2} \,.$$

The result matches the idea that the work required is the same as the work to move the whole amount of water, volume $w^2 h$, from the bottom to the *average* height within the tank, $h/2$.



Exercise 3.6.10

Suppose we are to fill a cylindrical tank with base of radius R and height h with water from the bottom. What is the work required?

Exercise 3.6.11

What if the horizontal cross-sections of the tank have arbitrary (but identical) shape?

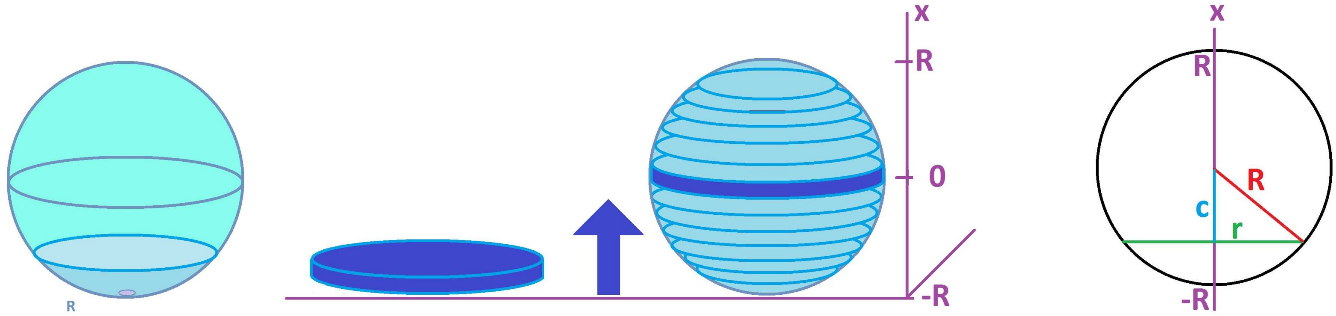
Exercise 3.6.12

Suppose a chain of weight M and length h is to be pulled all the way up from the ground? What is the work required?

In the examples above, the work is repetitive. What if the cross-section varies in shape and size?

Example 3.6.13: spherical tank

Suppose we are to fill a spherical tank of radius R with water from the bottom:



What is the work required?

We imagine that water appears at the bottom in thin slices and then each is delivered to the

appropriate height. They come from an augmented partition P of $[-R, R]$. The ith slice is a disk between the planes $x = x_{i-1}$ and $x = x_i$. Its thickness is $\Delta x_i = x_i - x_{i-1}$, radius r_i (to be found), and its weight is $\pi r_i^2 \cdot \Delta x_i$.

Now, the ith slice is delivered to location c_i (depicted negative), covering the interval $[-R, c_i]$. The displacement is, therefore, $R + c_i$ and the work to do so is

$$\pi r_i^2 \Delta x_i \cdot (R + c_i)\,.$$

Then the total work is approximated by the following:

$$\text{Work } \approx \sum_{i=1}^{n} \pi r_i^2 (R + c_i) \cdot \Delta x_i\,.$$

Let's find the radius of the slice. From the Pythagorean Theorem, we have:

$$r_i^2 = R^2 - c_i^2\,.$$

Then the above expression is the Riemann sum of the following integral:

$$\begin{aligned}
\text{work } &= \pi \int_{-R}^{R} (R^2 - x^2)(R + x)\, dx \\
&= \pi \int_{-R}^{R} \left(R^3 - x^2 R + R^2 x - x^3\right)\, dx \\
&= \pi \left(R^3 x - \frac{1}{3}x^3 R + R^2 \frac{1}{2} x^2 - \frac{1}{4}x^4\right)\Bigg|_{-R}^{R} \\
&= \pi \left(R^4 - \frac{1}{3}R^4 + R^4 \frac{1}{2} - \frac{1}{4}R^4\right) - \pi \left(-R^4 + \frac{1}{3}R^4 + R^4 \frac{1}{2} - \frac{1}{4}R^4\right) \\
&= \frac{4}{3}\pi R^4.
\end{aligned}$$

The result matches the idea that the work required is the same as the work to move the whole ball of water, volume $\frac{4}{3}\pi R^3$, so that its center of mass moves from $-R$ to 0.

Exercise 3.6.14

Suppose we are to fill a "paraboloid" tank acquired by rotating the graph of $y = x^2$ around the x-axis, which is vertical, from the bottom. What is the work required?

Exercise 3.6.15

What work is needed to pull all the way up a chain hanging down if it is 10 feet long and 20 pounds heavy?

Exercise 3.6.16

What is the meaning of the work required over an infinitely long trip?

Exercise 3.6.17

Show that the work required to move an object piece by piece, as above, is the same as the work required to move the whole object as if its total mass is concentrated at a single point, its center of mass.

3.7. The total and the average value of a function

What do these examples have in common?

A certain quantity, f, is "spread" around locations in space; for now, it is an interval within the x-axis. This quantity may be: length, area, density, velocity, force. When the quantity is constant within a segment of the interval, multiplying this value by the length of this piece, Δx, gives us a new but still familiar quantity:

quantity f	$f \cdot \Delta x$	$\Sigma f \cdot \Delta x$
length	area	total area
linear density	mass	total mass
flow rate	flux	total flux
force	work	total work

When the quantity f varies from segment to segment over the interval, it is represented by a function. When this change is incremental, the *total value* of f is the sum of the terms $f \cdot \Delta x$, i.e., the Riemann sum of the function f. When this change is continuous, the *total value* of f is approximated by this Riemann sum and, at the limit, it is the integral of f over $[a, b]$.

Now, the *average* value.

Recall that the *mean* (or the average) of a quantity given by n numbers $y_1, ..., y_n$ is defined to be the following:

$$\text{Mean } = \frac{y_1 + y_2 + ... + y_n}{n}\,.$$

How should we understand the mean of a quantity that is *continuously* spread over a line segment, say $[a, b]$? The numerator would have *infinitely many* terms!

Let's start with the idea of a *weighted average*. We assume that we have n *weights*, i.e., n positive numbers $m_1, ..., m_n$ with

$$m_1 + ... + m_n = 1\,.$$

Then for any given n numbers $y_1, ..., y_n$, we define their weighted average as follows:

$$\text{Weighted average } = m_1 y_1 + m_2 y_2 + ... + m_n y_n = \sum_{i=1}^{n} m_i y_i\,.$$

Exercise 3.7.1

Show that the mean is the weighted average with $m_i = 1/n$ for all i.

Example 3.7.2: scores

The weighted average may appear when one computes the total score in a class after several assignments of different *weights*. For example, this may be the grade breakdown:

- participation: 20%
- quizzes: 30%
- midterm: 20%
- final exam: 30%

Then the total score is the following weighted average of the five scores:

$$\text{TOTAL } = .20 \times P + .30 \times Q + .20 \times M + .30 \times F\,.$$

Example 3.7.3: center of mass?

Recall that if c is the coordinate of the center of mass of a system of weights with locations at y_i and relative weights m_i, then
$$c = \sum_i m_i y_i \,.$$
Therefore, the weighted average is the same, in this case, as the center of mass of the system.

The new setup is as follows:

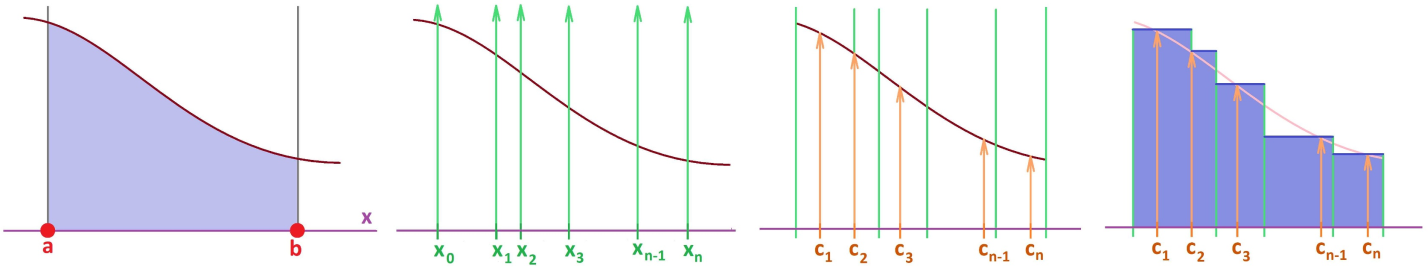
- ▶ A substance is uniformly distributed over a segment of the line.

Then, the weight of the segment is proportional to its width.

We are, then, justified to use these widths as substitutes for the *weights* in the weighted average. This is the main idea:

- ▶ Each weight m_i is the *relative length of the interval* where y_i of the quantity is located.

We start with an augmented partition P of the interval:
$$a = x_0 \le c_1 \le x_1 \le ... \le c_n \le x_n = b$$
Then we write the relative lengths:
$$m_i = \frac{\Delta x_i}{b-a} \,.$$
Let's substitute:
$$\text{Weighted average } = \sum_{i=1}^{n} \frac{\Delta x_i}{b-a} y_i = \frac{1}{b-a} \sum_{i=1}^{n} y_i \cdot \Delta x_i \,.$$
Furthermore, if these numbers are given by a function defined at the secondary nodes of the partition,
$$f(c_i) = y_i \,,$$
then we have:
$$\text{Weighted average } = \frac{1}{b-a} \sum_{i=1}^{n} f(c_i) \cdot \Delta x_i \,.$$



This sum is the Riemann sum of this function.

Definition 3.7.4: average value of function

The *average value* of a function f defined at the secondary nodes of a partition of an interval $[a, b]$ is denoted and defined as follows:
$$\bar{f} = \frac{1}{b-a} \Sigma\, f \cdot \Delta x \,.$$

It is, in other words, the total value of f per unit of length.

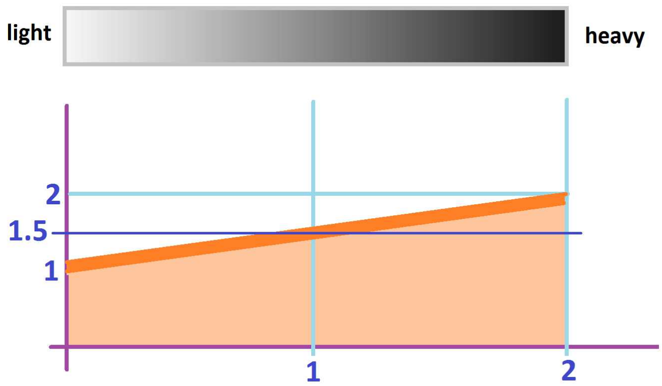
Warning!

The average mass of a system of weights is not the same as the average *location* of the weights.

What if the function doesn't change incrementally but "continuously"?

Example 3.7.5: linear dependence

Suppose the function f is linear, from 1 to 2 over the interval $[0, 2]$. Then the meaning of the average value is clear; it is 1.5.



Where does it comes from? The area of the triangle under the graph of f, i.e., 3, which is the integral of f over $[0, 2]$, divided by its length, 2.

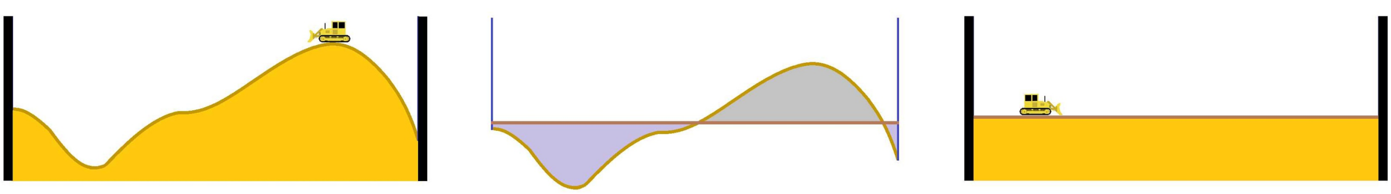
Then we think of the fraction above as an approximation of the average. This analysis justifies the following definition:

Definition 3.7.6: average value of function

The *average value* of an integrable function f over interval $[a, b]$ is denoted and defined as follows:

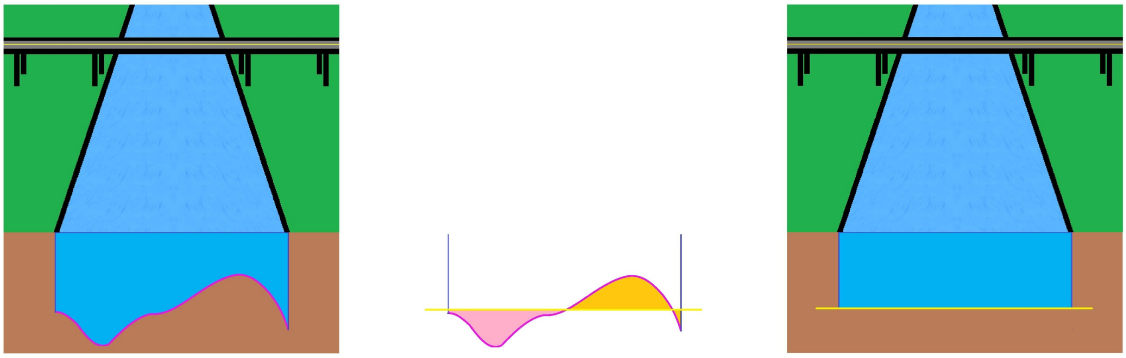
$$\bar{f} = \frac{1}{b-a}\int_a^b f\,dx\,.$$

To illustrate, consider how one levels an uneven surface of sand:



The amount of sand is the same.

The average depth of a canal is another interpretation:



Both canals have the same amount of water.

Thus, by "averaging" we mean replacing any function, $y = f(x)$, with a *constant function*, $y = \bar{f}$, chosen so that the two have the same integral:

$$\bar{f} \cdot (b - a) = \int_a^b f \, dx \, .$$

Exercise 3.7.7

Prove the above statement.

Theorem 3.7.8: Properties of Average

Over a given interval, we have:

- *The average of the sum is the sum of the averages:*

$$\overline{f + g} = \bar{f} + \bar{g} \, .$$

- *The constant multiple of the average is the average of the constant multiple:*

$$\overline{cf} = c\bar{f} \, .$$

Exercise 3.7.9

Prove the theorem.

Exercise 3.7.10

What can you say about the average of (a) an odd function over $[-r, r]$, (b) an even function over $[-r, r]$, (c) a periodic function?

We can rewrite the table with which we started the section:

f	$\int_a^b f \, dx$	$\frac{1}{b-a} \int_a^b f \, dx$
length	total area	average length
linear density	total mass	average linear density
flux	total flux	average flux
force	total work	average force

All have been introduced in the chapter except for the first item that comes from Volume 2.

Exercise 3.7.11

What is the meaning of the average of a function defined on an infinite interval?

3.8. Numerical integration

To apply the integral formulas presented in this chapter, we need to evaluate those integrals.

This is an ideal outcome:

$$\text{Area } = \int_0^2 x^2\,dx = \frac{x^3}{3}\bigg|_0^1 = \frac{1^3}{3} - \frac{0^3}{3} = \frac{1}{3}\,.$$

We have an *exact number*.

However, such an outcome is an exception, not a rule! Some integrals do not produce familiar functions so that we can just plug in the two values. Conversely, some of the functions have been defined as integrals only, such as:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int e^{-x^2}\,dx\,.$$

There is no other formula!

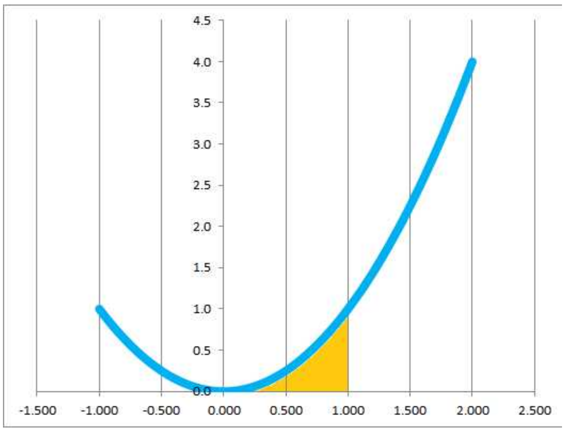
What do we do? The answer is in the definition of the Riemann integral. It is defined via the Riemann sums of the function and these sums serve as its approximations.

We will assume below that all functions are *integrable*, which means that any choice of the secondary nodes for the Riemann sums is equally valid.

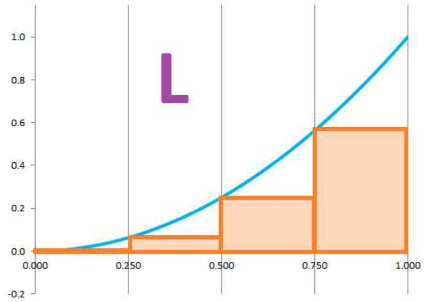
Example 3.8.1: partition schemes

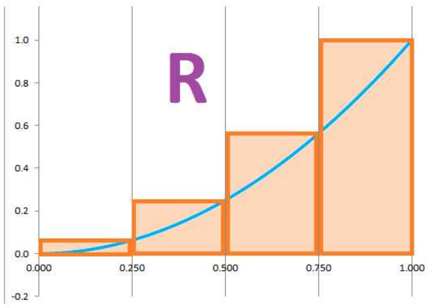
Let's review the ways we estimate this integral of $f(x) = x^2$ over $[0,1]$:

$$\int_0^1 f\,dx\,.$$



We choose the number of intervals to be $n = 4$ with equal intervals of length $h = 1/4$. Then we choose, as the secondary nodes, the left-end or the right-end of each interval:





At those points, the function is evaluated. This is the computation of the left-end Riemann sum L_4:

	$\bullet$	$--$	$\vert$	$--$	$\vert$	$--$	$\vert$	$--$	$\bullet$	
x	0		$1/4$		$1/2$		$3/4$		1	
x^2	0		$1/16$		$1/4$		$9/16$		1	
L_4	$0\cdot 1/4$	$+$	$1/16\cdot 1/4$	$+$	$1/4\cdot 1/4$	$+$	$9/16\cdot 1/4$			≈ 0.22

We, furthermore, realize that we are computing the Riemann sum *function* for this augmented partition. Its four values are shown at bottom of the table.

Warning!

It is better to avoid "approximating with approximations" and replace the last number with its exact value:

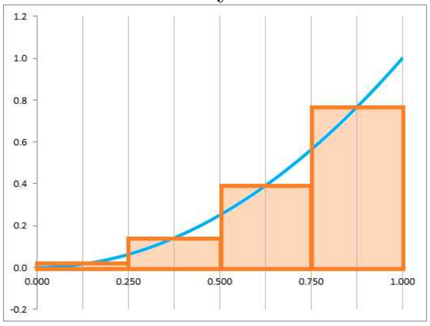
$$L_4 = \frac{7}{32} = .21875\,.$$

Exercise 3.8.2

Create a table of values for the Riemann sum function for the right ends.

Example 3.8.3: mid-point Riemann sums

We can also choose the mid-points as the secondary nodes:



This is the computation of the mid-point Riemann sum M_4 for the same integral:

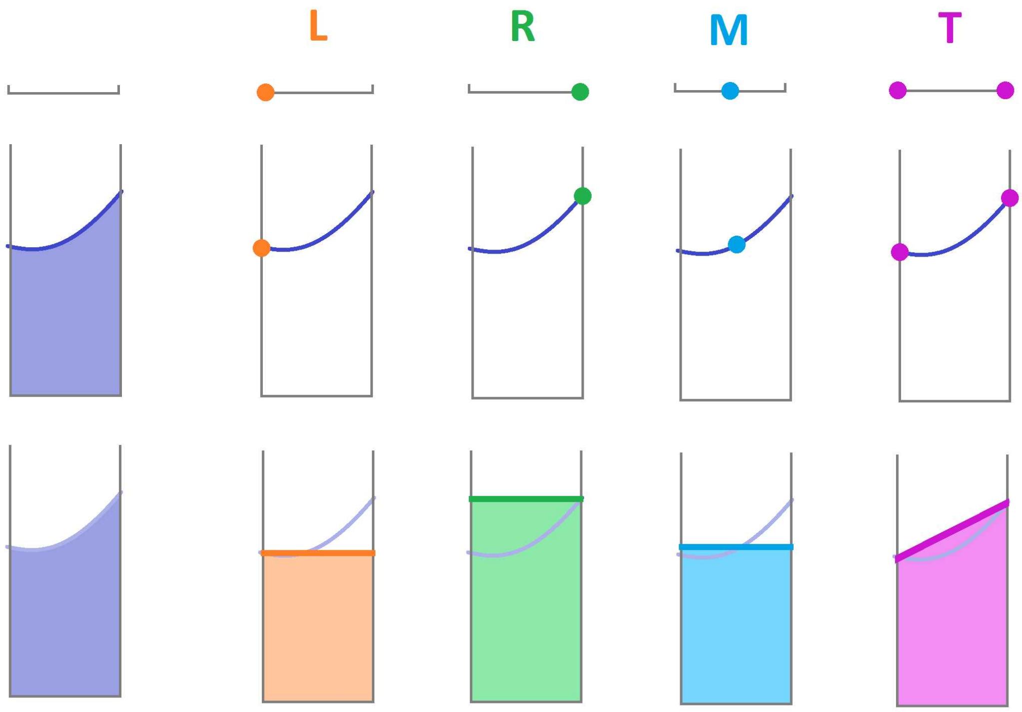
	$\bullet$	$--$	$\vert$	$--$	$\vert$	$--$	$\vert$	$--$	$\bullet$
x		$1/8$		$3/8$		$5/8$		$7/8$	
$f(x)=x^2$		$(1/8)^2$		$(3/8)^2$		$(5/8)^2$		$(7/8)^2$	
M_4		$(1/8)^2\cdot 1/4$	$+$	$(3/8)^2\cdot 1/4$	$+$	$(5/8)^2\cdot 1/4$	$+$	$(7/8)^2\cdot 1/4$	$=0.328125$

It is much closer than L_4 to the true value of the integral, $1/3$.

Exercise 3.8.4

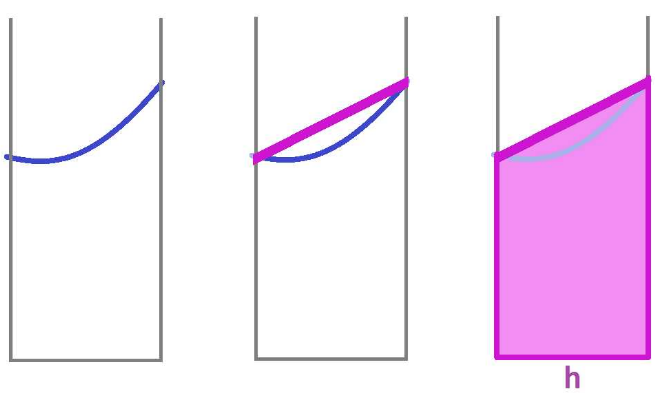
We have previously used a spreadsheet to calculate the Riemann sum function for L_n. Create a spreadsheet to automate computations of the Riemann sum function for R_n and M_n.

This is what all Riemann sums have in common: We choose a point on the graph and then approximates its piece with a *horizontal* segment. The three main choices are shown below:



What if we choose *two points* – at the end and the beginning of the interval – and approximate this piece of the graph with a *sloped* line? It is, in fact, the familiar secant line! This third way to approximate the area is shown on the far right.

Instead of a rectangle, we have a *trapezoid*. Its area is the average of the lengths of the two bases (vertical) multiplied by the height (horizontal):



Then the area of the trapezoid over the interval $[x_{k-1}, x_k]$ is equal to

$$\frac{f(x_{k-1}) + f(x_k)}{2} h\,.$$

The sum of all n of these is called the *trapezoid approximation of the integral* and denoted by T_n.

Exercise 3.8.5

Show that T_n is the average of L_n and R_n.

Example 3.8.6: T_n

Let's compute sum T_4 for the same integral. We use the same data and then add the following terms:

$$f(x_{k-1})h + f(x_k)h$$

for each interval:

	$\bullet$	$--$	$\vert$	$--$	$\vert$	$--$	$\vert$	$--$	$\bullet$	
x	0		$1/4$		$1/2$		$3/4$		1	
x^2	0		$1/16$		$1/4$		$9/16$		1	
	$0 \cdot 1/4$	$+$	$1/16 \cdot 1/4$							≈ 0.016
			$1/16 \cdot 1/4$	$+$	$1/4 \cdot 1/4$					≈ 0.079
					$1/4 \cdot 1/4$	$+$	$9/16 \cdot 1/4$			≈ 0.203
							$9/16 \cdot 1/4$	$+$	$1 \cdot 1/4$	≈ 0.391
									sum	≈ 0.689
T_4									half	≈ 0.345

Warning!

The result is *not* a Riemann sum.

Exercise 3.8.7

Create a spreadsheet to automate computations of T_n.

These are the formulas for the four approximations:

$$L_n = \sum_{i=1}^{n} f\big(a + (i-1)h\big)h$$

$$R_n = \sum_{i=1}^{n} f\big(a + ih\big)h$$

$$M_n = \sum_{i=1}^{n} f\big(a + (i-1)h + h/2\big)h$$

$$T_n = \sum_{i=1}^{n} \frac{1}{2}\Big[f\big(a + (i-1)h\big) + f\big(a + ih\big)\Big]h$$

The expressions approximate the integral in the following sense.

Theorem 3.8.8: Convergence of Riemann Sums

If f is integrable on $[a,b]$, then the sequences of the Riemann sums and the trapezoid sum converge to the Riemann integral of f:

$$L_n,\ R_n,\ M_n,\ T_n \ \to \int_a^b f\,dx \quad \text{as} \quad n \to \infty\,.$$

Only the last part needs proof.

Exercise 3.8.9

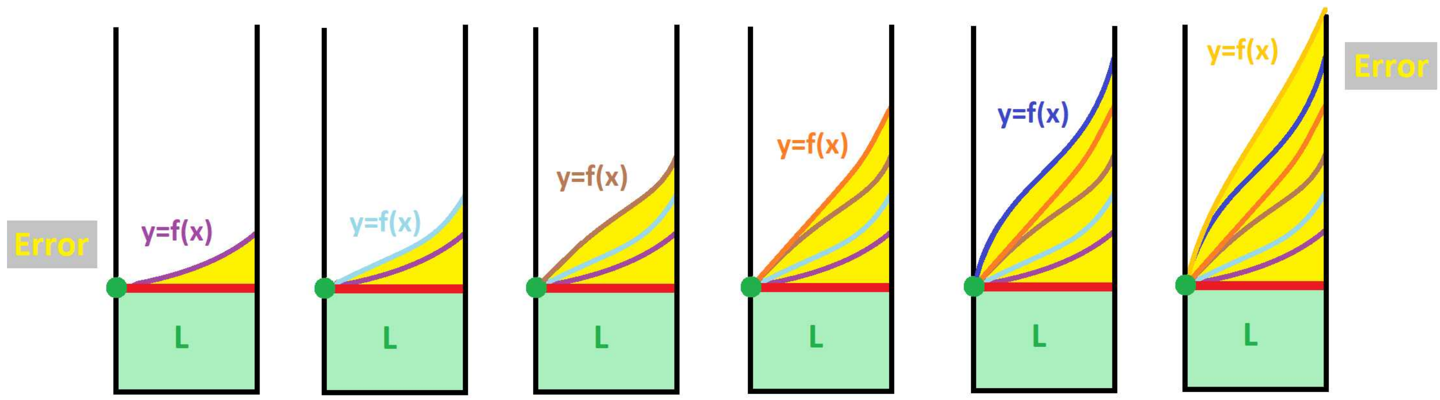
Prove the missing part. Hint: The Squeeze Theorem.

How well do these four perform?

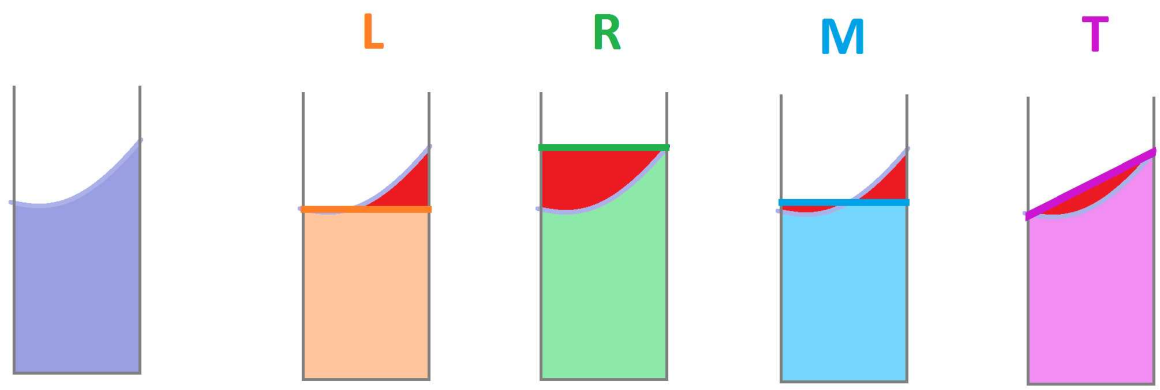
- Question: For a given n, how close are we to the true value of the integral?
- Answer: We don't know, and we can't know, without some *a priori* knowledge.

If we knew the value of the integral, we wouldn't need to approximate it!

Various behaviors of f are shown along with the error (yellow) of L_n:



Since all four approximations are specific areas, the errors are also seen as certain areas:



The simplest *a priori* knowledge about a function is its monotonicity.

Theorem 3.8.10: Riemann Sums of Monotone Functions

1. *If f is increasing on $[a, b]$, the left-end Riemann sum underestimates the integral, while the right-end sum overestimates it:*

$$f \nearrow \implies L_n \leq \int_a^b f\,dx \leq R_n\,.$$

2. *When the function is decreasing, the inequalities are reversed:*

$$f \searrow \implies L_n \geq \int_a^b f\,dx \geq R_n\,.$$

In either case, the true value of the integral I lies between R_n and L_n. In other words, we have:

Corollary 3.8.11: Interval for Integral of Monotone Function

If a function f is monotone, then

$$\int_a^b f\,dx \;\; \text{lies within either} \;\; [L_n, R_n] \;\; \text{or} \;\; [R_n, L_n]\,.$$

The statement is the absolute, not approximate, truth!

While the last result relies on monotonicity (and the derivative of f), for the other two approximations, a similar result relies on concavity (and the second derivative of f).

Theorem 3.8.12: Riemann Sums of Convex Functions

1. *If f is concave down on $[a,b]$, the trapezoid sum underestimates the integral:*
$$f\ \frown\ \Longrightarrow\ T_n \le \int_a^b f\, dx\,.$$
2. *Meanwhile, when the function is concave up, the inequality is reversed:*
$$f\ \smile\ \Longrightarrow\ T_n \ge \int_a^b f\, dx\,.$$

We thus have discovered how these approximations err in different *directions* under different circumstances. However, the true measure of the quality of the approximation is the actual difference, i.e., the *distance* from the integral:

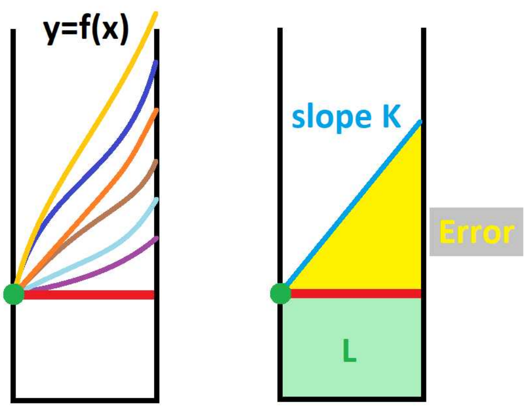
$$\text{Error } = \Bigg|\text{ Integral } - \text{ Approximation }\Bigg|$$

Since we don't know the value of the integral, we don't know the value of the error; we can only estimate it. With this estimate, we can establish that we haven't deviated from the truth too far.

Exercise 3.8.13

Estimate the error when the function is monotone.

Let's take a look at the left-end approximation on a single interval. Suppose the only value that matters, $f(x_k)$, is known. Beyond that, the function can exhibit a variety of behaviors, including fast *growth*. The faster f grows past x_k, the larger is the error of L_n. As the rate of this growth is limitless, so is our error (left):



Can we control the size of the error? Yes, if we are aware of – *a priori* – the limit on the rate of growth of f, i.e., its derivative. If the derivative is less than some number K, then the slope of the graph of f is less than K and, therefore, the graph will have to stay under the line with slope K (right). This line is the worst-case scenario.

Note that such a restriction is expected to be possible when the derivative f' is continuous according to the *Extreme Value Theorem* (Volume 2).

Below, we provide a bound for the error:

Theorem 3.8.14: Error Bound of Riemann Sums I

Suppose a function f is differentiable and for all x in $[a, b]$ we have:

$$|f'(x)| \leq K_1 \,,$$

for some number K_1. Then

$$\left| S_n - \int_a^b f\, dx \right| \leq \frac{K_1(b-a)^2}{2n} \,,$$

whether S_n is the left $L_n(f)$ or the right $R_n(f)$ Riemann sum of f.

Proof.

If we have an estimate for the derivative,

$$f'(x) \leq K \,,$$

on the whole domain of integration $[a, b]$, we have an estimate for the error on each interval first:

$$\left| f(x_k)\Delta x - \int_{x_{k-1}}^{x_k} f\, dx \right| \leq \frac{1}{2}(K\Delta x) \cdot \Delta x = \frac{1}{2} K \Delta x^2 \,,$$

as the area of this triangle. Then we compute an estimate for the error of the left-end Riemann sum over the whole interval:

$$\begin{aligned}
E_n &= \left| L_n - \int_a^b f\, dx \right| \\
&= \left| \sum_{k=0}^{n-1} f(x_k)\Delta x - \sum_{k=0}^{n-1} \int_{x_{k-1}}^{x_k} f\, dx \right| \\
&= \left| \sum_{k=0}^{n-1} \left(f(x_k)\Delta x - \int_{x_{k-1}}^{x_k} f\, dx \right) \right| \\
&\leq \sum_{k=0}^{n-1} \left| f(x_k)\Delta x - \int_{x_{k-1}}^{x_k} f\, dx \right| && \text{By the Triangle Inequality in Chapter 1PC-2.} \\
&\leq \sum_{k=0}^{n-1} \frac{1}{2} K \Delta x^2 \\
&= \sum_{k=0}^{n-1} \frac{1}{2} K \left(\frac{b-a}{n} \right)^2 \\
&= \sum_{k=0}^{n-1} \frac{1}{2} K \frac{(b-a)^2}{n^2} \\
&= \frac{1}{2} K \frac{(b-a)^2}{n} \,.
\end{aligned}$$

So, an *a priori* bound on the derivative gives an *a priori* bound on the error.

If now we need to say something specific about the unknown value of the integral, we can:

Corollary 3.8.15: Interval for Riemann Integral

Suppose a function f is differentiable and for all x in $[a, b]$ we have:

$$|f'(x)| \leq K_1 \,,$$

for some number K_1. Let

$$E_n = \frac{K_1(b-a)^2}{2n}.$$

Then, the integral

$$\int_a^b f\,dx$$

lies within the following interval:

$$\left[S_n - E_n, S_n + E_n\right],$$

whether S_n is the left $L_n(f)$ or the right $R_n(f)$ Riemann sum of f.

Example 3.8.16: error bound of Riemann sums

Let's test this theorem on the integral

$$\int_0^1 x^2\,dx = 1/3,$$

with $L_4 = 0.22$ computed previously. First, we find the derivative:

$$f(x) = x^2 \implies f'(x) = 2x.$$

Then, we choose, of course,

$$K_1 = 2.$$

Next,

$$E_4 = \frac{2(1-0)^2}{2\cdot 4} = \frac{2}{8} = 0.25.$$

Then the integral's value should be within the interval:

$$[L_4 - E_4,\ L_4 + E_4] = [0.22 - 0.25,\ 0.22 + 0.25] = [-0.03,\ 0.45].$$

A very crude but correct estimate!

We can apply the theorem again, to the right-end approximation, resulting in an interval of the same size but centered around R_4:

$$[R_4 - E_4,\ R_4 + E_4] = [0.47 - 0.25,\ 0.47 + 0.25] = [0.22,\ 0.72].$$

Furthermore, since the value I of the integral belongs to both intervals, it belongs to their intersection:

$$[-0.03,\ 0.45] \cap [0.22,\ 0.72] = [0.22,\ 0.45].$$

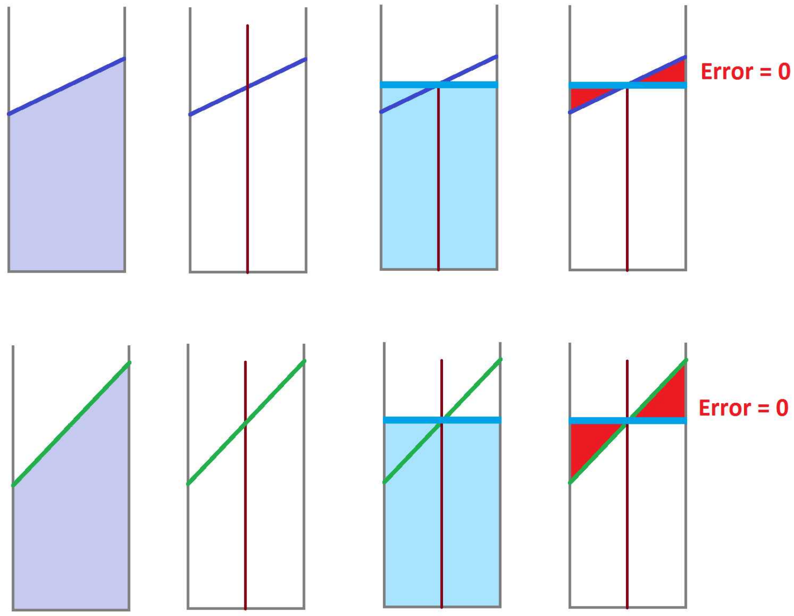
Similarly, we have I within

$$[L_{10} - E_{10},\ L_{10} + E_{10}] = [0.29 - 0.1,\ 0.29 + 0.1] = [0.19,\ 0.39].$$

Exercise 3.8.17

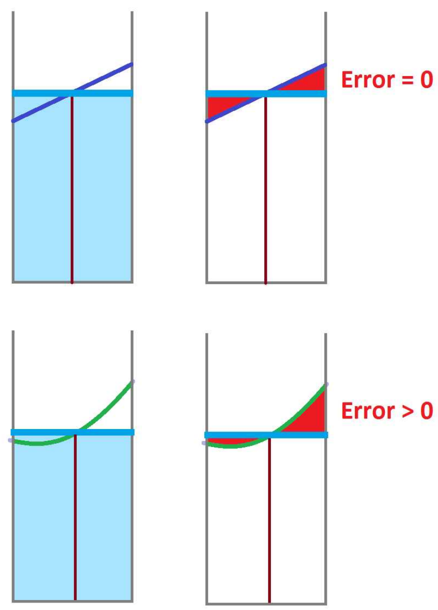
Find the interval for the above integral using R_{10}.

For contrast, let's take a look at the mid-point approximation. First suppose that f is linear:



Even though the slopes are different, the error is the same, zero. It appears then that the *derivative* doesn't matter!

Let's now add concavity:



The error isn't zero as in the former case. This observation suggests that the error is "created" by the *second derivative* of f.

Exercise 3.8.18

What difference does it make if f is concave down instead of up?

The idea of the last theorem was to use a bound for the *derivative* to make sure that the function doesn't deviate too far from its *constant* approximation (on each interval). Similarly, the idea of the next theorem is to use a bound for the *second derivative* to make sure that the function doesn't deviate too far from its *linear* approximation. We accept the result without proof.

Theorem 3.8.19: Error Bound of Riemann Sums II

Suppose for all x in $[a, b]$, we have

$$|f''(x)| \leq K_2 \,,$$

for some real K_2. *Then*

$$\left| M_n(f) - \int_a^b f\,dx \right| \leq \frac{K_2(b-a)^3}{24n^2}\,,$$

and

$$\left| T_n(f) - \int_a^b f\,dx \right| \leq \frac{K_2(b-a)^3}{12n^2}\,.$$

So, an *a priori* bound on the second derivative gives an *a priori* bound on the error.

Exercise 3.8.20

Suggest a similar theorem for L_n and R_n. Hint: What is the worst-case scenario?

Thus, the true value of the integral lies within this interval:

$$[M_n - E_n, M_n + E_n]\,,$$

where

$$E_n = \frac{K_2(b-a)^3}{24n^2}\,.$$

Example 3.8.21: error bound of Riemann sums, continued

Let's confirm this result for

$$\int_0^1 x^2\,dx = 1/3$$

and $M_4 = 0.328125$ computed previously. First, we find the derivatives:

$$f(x) = x^2 \implies f'(x) = 2x \implies f''(x) = 2\,.$$

Then, we choose, of course,

$$K_2 = 2\,.$$

Next,

$$E_4 = \frac{2(1-0)^3}{24\cdot 4^2} = \frac{2}{24\cdot 16} = 0.0052083333...$$

Then the integral's value should be within the interval:

$$[M_4 - E_4, M_4 + E_4] = [0.328125 - 0.0052083..., 0.328125 + 0.0052083...] = [0.329166..., 0.333333...]\,.$$

It happens to be exactly the right end of the interval. The reason is that K_2 isn't an estimate but the exact value of the second derivative.

Example 3.8.22: more complex error bound

A more complex example is:

$$\int_0^1 x^3\,dx = 1/4\,.$$

First, the estimate of the integral with $n = 4$:

$$M_4 = (1/8)^3\cdot 1/4 + (3/8)^3\cdot 1/4 + (5/8)^3\cdot 1/4 + (7/8)^3\cdot 1/4 = 0.2421875\,.$$

Then, we find the derivatives:

$$f(x) = x^3 \implies f'(x) = 3x^2 \implies f''(x) = 6x\,.$$

We need K_2 to satisfy:

$$K_2 \geq |f''(x)| = 6x,\ \text{for all } 0 \leq x \leq 1\,.$$

The choice is then obvious:

$$K_2 = 6\,.$$

Next, the error bound:

$$E_4 = \frac{K_2(b-a)^3}{24n^2} = \frac{6(1-0)^3}{24\cdot 4^2} = \frac{6}{24\cdot 16} = 0.015625\,.$$

Then the integral's value should be within the interval:

$$[M_4 - E_4,\ M_4 + E_4] = [0.242 - .016,\ 0.242 + 0.016] = [0.226,\ 0.258]\,.$$

It is.

Note that the existence of K_2 is guaranteed by the *Extreme Value Theorem* provided f'' is continuous.

Example 3.8.23: how to guarantee accuracy

At the next, more practical, level, we are asked to estimate an integral with a *given* accuracy. For example, suppose we need to know within 0.1 the value of the integral:

$$\int_0^1 x^3\,dx\,.$$

Then the answer above applies as $E = 0.015625 < 0.1$.

What if the accuracy needs to be 0.01? Then $n = 4$ is too small! Let's try $n = 5$. We have:

$$E_5 = \frac{K_2(b-a)^3}{24\cdot 5^2} = \frac{6(1-0)^3}{24\cdot 5^2} = \frac{6}{24\cdot 25} = 0.01\,.$$

Furthermore, we observe that in order to ensure that the error is less than some $\varepsilon > 0$, we simply need to find n that satisfies:

$$\frac{6(1-0)^3}{24\cdot n^2} \le \varepsilon\,.$$

In general, we are solving the inequality:

$$E_n = \frac{K_2(b-a)^3}{24n^2} \le \varepsilon\,.$$

Corollary 3.8.24: Estimation of Error of Numerical Integration

Suppose for all x in $[a,b]$ we have

$$|f''(x)| \le K_2\,,$$

for some real K_2. Then, for any given $\varepsilon > 0$, the value of the integral

$$\int_a^b f\,dx$$

lies within ε from M_n provided

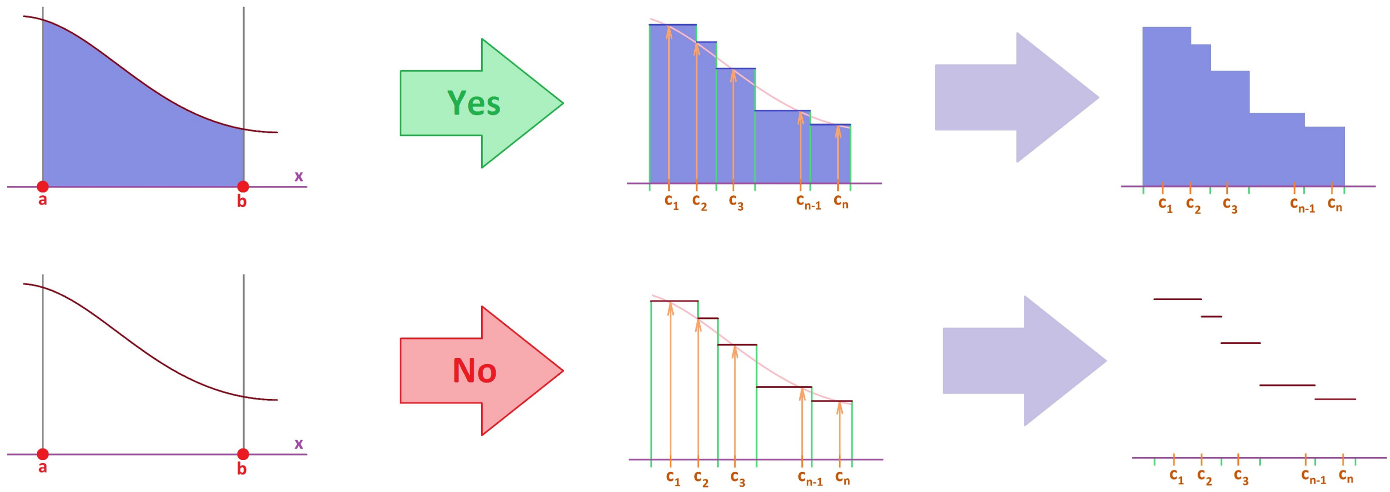
$$n \ge \sqrt{\frac{K_2(b-a)^3}{24\varepsilon}}\,.$$

Exercise 3.8.25

Create a spreadsheet to automate these computations.

3.9. *Lengths of curves*

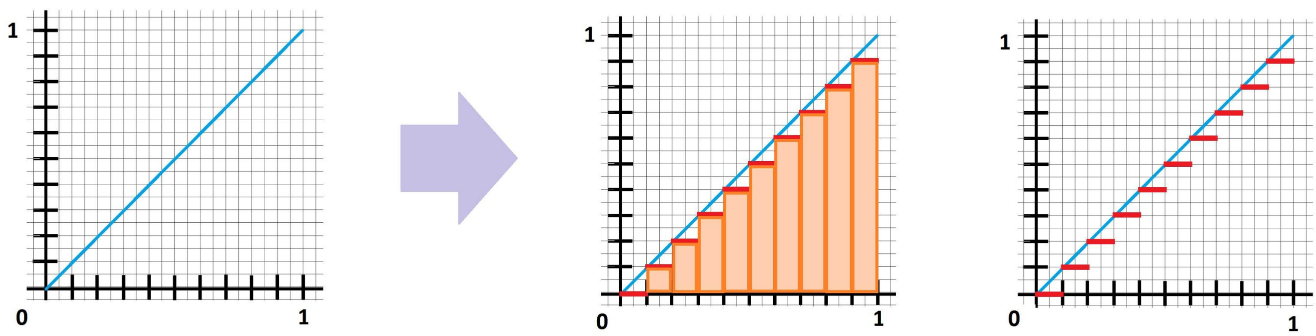
We have successfully used the Riemann sum construction to approximate and, at the limit, compute the *areas* under the graphs of functions. It would be, however, a grave mistake to think that the step function produced by this construction can serve as an approximation of the function itself:



The reason is revealed when we watch how spectacularly this idea fails when applied to computing the *lengths* of curves.

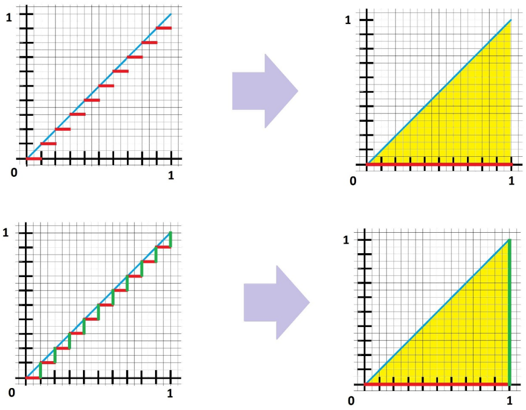
Example 3.9.1: straight line

Let's consider a very simple case of $y = f(x) = x$ over $[0, 1]$. The approximation of the curve with the horizontal segments looks just as good as the approximation of the area under the graph:



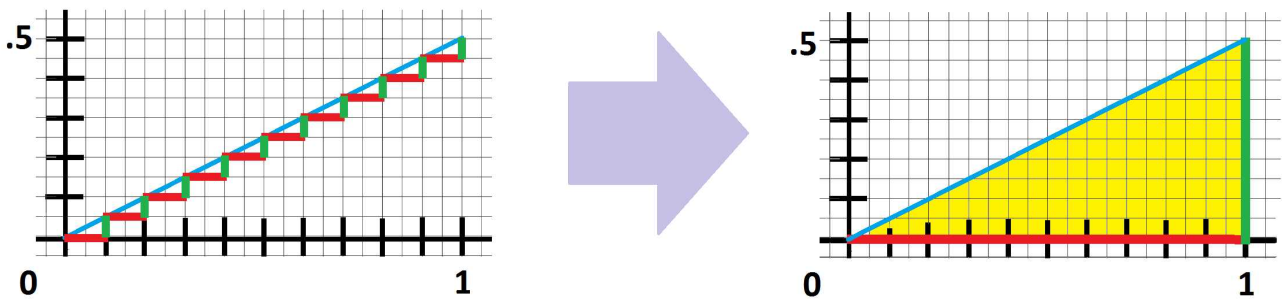
The result is illustrated for a partition with $n = 10$ intervals of equal length and the left ends as secondary nodes.

A problem appears when we look at the actual numbers. The length of the original graph is $\sqrt{2}$ by the *Pythagorean Theorem*. Meanwhile, the total length of the horizontal segments that make up the graph of the resulting step function is 1; it's simply the bottom of the big triangle. Too low!



One may try to fix the problem by adding the vertical segments to our estimate of the length of the diagonal. Then, the estimate becomes 2; it's simply the sum of the other two sides of the big triangle. Too high! It is important that the numbers won't change even if we start to refine the partition. In contrast, the approximation of the area of the triangle, L_n, is getting better as we increase n.

To understand the reason for this discrepancy, let's consider the line $y = g(x) = x/2$. Its actual length is $\sqrt{1^2 + (1/2)^2} \approx 1.19$ by the *Pythagorean Theorem*. The approximation with horizontal segments is still equal to 1 and the one with both vertical and horizontal segments is 1.5. The estimates are still off but they are closer to the truth!



What explains the difference? The *slope*. To confirm this idea, just take the line with *zero* slope. Then the estimate is equal to its actual length!

In fact, the case of a linear f is very simple:

$$\text{Length } = \sqrt{\text{ base }^2 + \text{ height }^2} = \sqrt{\text{ base }^2 + (\text{ base } \cdot \text{ slope })^2}\,.$$

Exercise 3.9.2

Show that the conclusions remain valid no matter what augmented partition of $[0,1]$ we choose.

Exercise 3.9.3

Show that the length of the graph of a step function over any partition of $[a,b]$ is $b-a$.

The lesson is that the approximations of the length of the graph – unlike the one for the area under the graph – should depend on the *derivative* of the function.

Example 3.9.4: length of circle

But first, let's compute the length of the circle as the graph of a function. We compute the length of the upper half of the unit circle by first representing it as the graph of a simple function:

$$f(x) = \sqrt{1-x^2}\,.$$

The idea is simple: Place points on the curve, connect them consecutively by edges, and then approximate the curve with a continuous curve made of these edges.

We have a list of the values of x in the first column:

$$x_0, x_1, ..., x_n\,,$$

and the list of corresponding values of y in the next column:

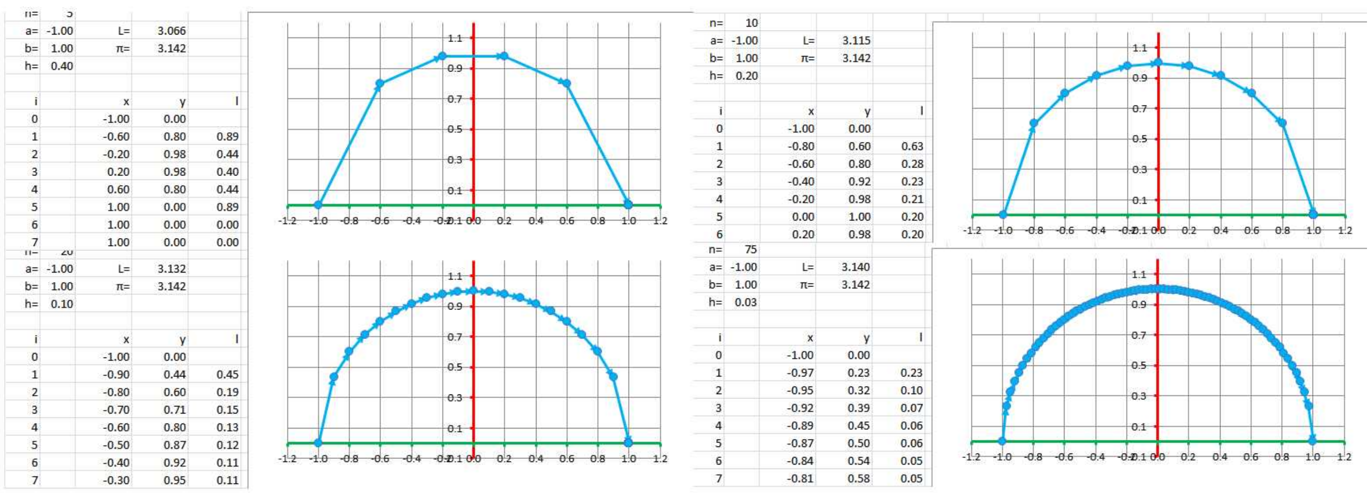
$$y_0 = f(x_0),\ y_1 = f(x_1), ...,\ y_n = f(x_n)\,.$$

In the third column, we compute the lengths of the segments via the *Distance Formula*:

$$l_k = \sqrt{(x_{k+1} - x_k)^2 + (y_{k+1} - y_k)^2}\,.$$

We use the formula:

```
=SQRT((RC[-2]-R[-1]C[-2])^2+(RC[-1]-R[-1]C[-1])^2)
```



n=	5			
a=	-1.00	L=	3.066	
b=	1.00	π=	3.142	
h=	0.40			
i		x	y	l
0		-1.00	0.00	
1		-0.60	0.80	0.89
2		-0.20	0.98	0.44
3		0.20	0.98	0.40
4		0.60	0.80	0.44
5		1.00	0.00	0.89
6		1.00	0.00	0.00
7		1.00	0.00	0.00

n=	10			
a=	-1.00	L=	3.115	
b=	1.00	π=	3.142	
h=	0.20			
i		x	y	l
0		-1.00	0.00	
1		-0.80	0.60	0.63
2		-0.60	0.80	0.28
3		-0.40	0.92	0.23
4		-0.20	0.98	0.21
5		0.00	1.00	0.20
6		0.20	0.98	0.20

n=	20			
a=	-1.00	L=	3.132	
b=	1.00	π=	3.142	
h=	0.10			
i		x	y	l
0		-1.00	0.00	
1		-0.90	0.44	0.45
2		-0.80	0.60	0.19
3		-0.70	0.71	0.15
4		-0.60	0.80	0.13
5		-0.50	0.87	0.12
6		-0.40	0.92	0.11
7		-0.30	0.95	0.11

n=	75			
a=	-1.00	L=	3.140	
b=	1.00	π=	3.142	
h=	0.03			
i		x	y	l
0		-1.00	0.00	
1		-0.97	0.23	0.23
2		-0.95	0.32	0.10
3		-0.92	0.39	0.07
4		-0.89	0.45	0.06
5		-0.87	0.50	0.06
6		-0.84	0.54	0.05
7		-0.81	0.58	0.05

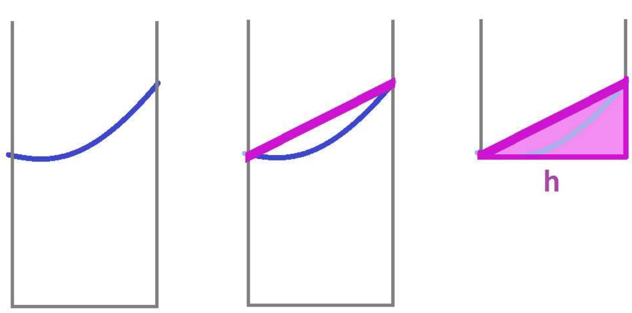
As we increase the number of segments, n, the result that we know to be correct, π, is being approached. We will see in Chapter 4 a better way to represent curves and especially the circle.

We, just as always, start with a discrete situation.

We simply have a sequence of points on the plane. Such a sequence is seen as a "curve" if we proceed from point to point along a straight line. The lengths of these segments are found by the *Distance Formula*, just as in the above example. It is this simple!

Now, something more specific. What if these points form the graph of a function $y = f(x)$ defined at the nodes of a partition of $[a, b]$?

This is what happens to each interval $[x_{k-1}, x_k],\ k = 1, 2, ..., n$ of the partition. The graph of f goes (jumps) from $(x_{k-1}, f(x_{k-1}))$ to $(x_k, f(x_k))$. We then construct a sloped segment between these two points:



It is the *secant line*! A right triangle is formed by these two segments:

- horizontal $[x_{k-1}, x_k]$, and
- vertical from $f(x_{k-1})$ to $f(x_k)$, or vice versa.

The lengths of these sides are:

- horizontal (base, the run): $h = x_k - x_{k-1} = \Delta x_k$, and
- vertical (height, the rise): $|f(x_k) - f(x_{k-1})| = \Delta y_k$.

The length of the edge (the hypotenuse of the triangle) is then the following:

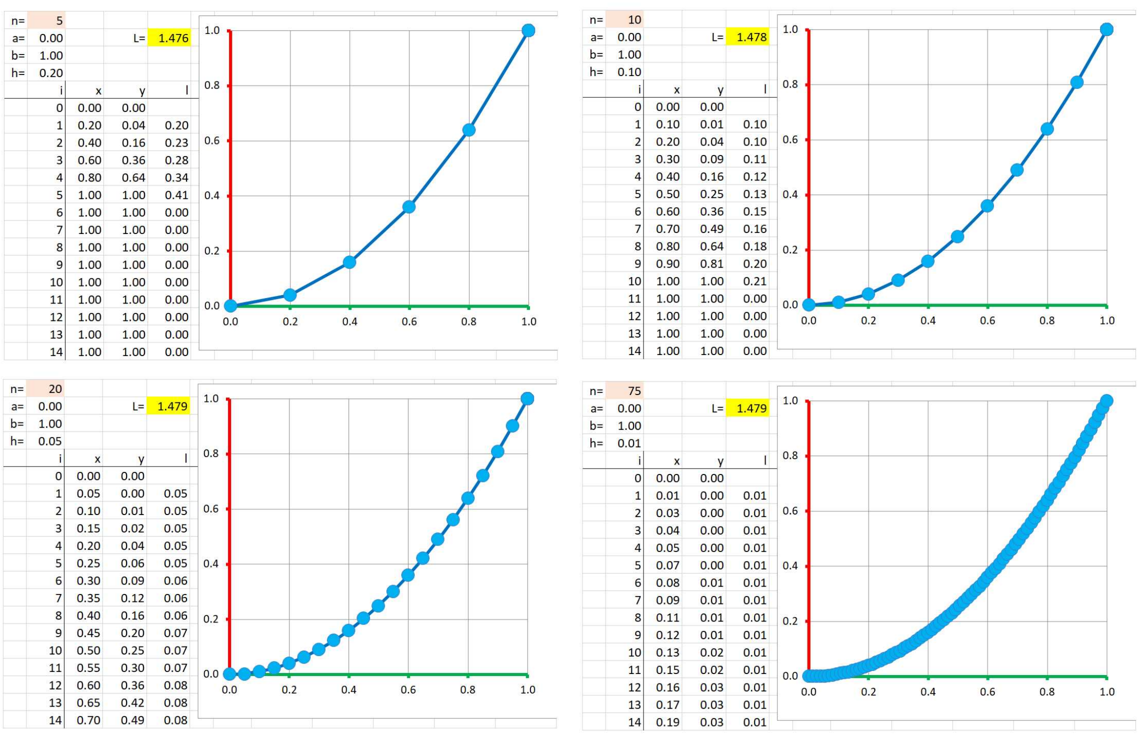
$$\sqrt{\Delta x_k^2 + \Delta y_k^2} = \sqrt{\Delta x_k^2 + (f(x_k) - f(x_{k-1}))^2}\,.$$

Thus, the full length of the trip along these points on the graph of f is equal to the following:

$$\boxed{\text{Total length } = \sum_{k=1}^{n} \sqrt{\Delta x_k^2 + \Delta y_k^2} = \sum_{k=1}^{n} \sqrt{\Delta x_k^2 + (f(x_k) - f(x_{k-1}))^2}}$$

Example 3.9.5: length of parabola

We approximate the length of a parabola $y = x^2$, $0 \le x \le 1$, below:



n=	5			
a=	0.00		L=	1.476
b=	1.00			
h=	0.20			
	i	x	y	l
	0	0.00	0.00	
	1	0.20	0.04	0.20
	2	0.40	0.16	0.23
	3	0.60	0.36	0.28
	4	0.80	0.64	0.34
	5	1.00	1.00	0.41
	6	1.00	1.00	0.00
	7	1.00	1.00	0.00
	8	1.00	1.00	0.00
	9	1.00	1.00	0.00
	10	1.00	1.00	0.00
	11	1.00	1.00	0.00
	12	1.00	1.00	0.00
	13	1.00	1.00	0.00
	14	1.00	1.00	0.00

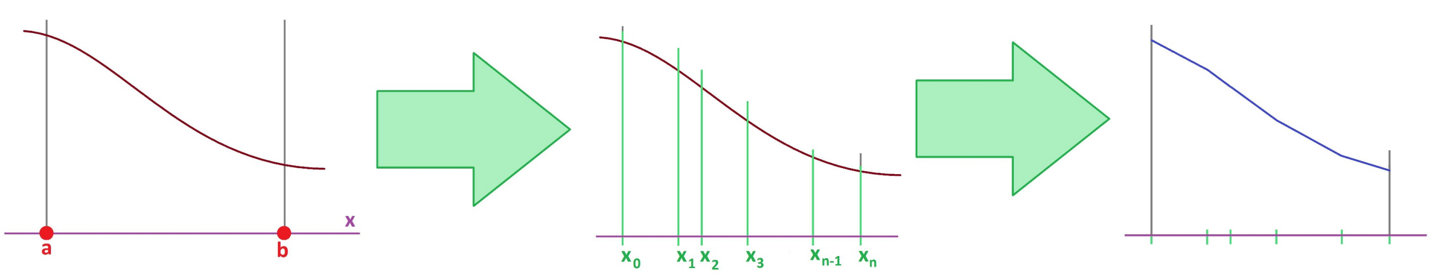
n=	10			
a=	0.00		L=	1.478
b=	1.00			
h=	0.10			
	i	x	y	l
	0	0.00	0.00	
	1	0.10	0.01	0.10
	2	0.20	0.04	0.10
	3	0.30	0.09	0.11
	4	0.40	0.16	0.12
	5	0.50	0.25	0.13
	6	0.60	0.36	0.15
	7	0.70	0.49	0.16
	8	0.80	0.64	0.18
	9	0.90	0.81	0.20
	10	1.00	1.00	0.21
	11	1.00	1.00	0.00
	12	1.00	1.00	0.00
	13	1.00	1.00	0.00
	14	1.00	1.00	0.00

n=	20			
a=	0.00		L=	1.479
b=	1.00			
h=	0.05			
	i	x	y	l
	0	0.00	0.00	
	1	0.05	0.00	0.05
	2	0.10	0.01	0.05
	3	0.15	0.02	0.05
	4	0.20	0.04	0.05
	5	0.25	0.06	0.05
	6	0.30	0.09	0.06
	7	0.35	0.12	0.06
	8	0.40	0.16	0.06
	9	0.45	0.20	0.07
	10	0.50	0.25	0.07
	11	0.55	0.30	0.07
	12	0.60	0.36	0.08
	13	0.65	0.42	0.08
	14	0.70	0.49	0.08

n=	75			
a=	0.00		L=	1.479
b=	1.00			
h=	0.01			
	i	x	y	l
	0	0.00	0.00	
	1	0.01	0.00	0.01
	2	0.03	0.00	0.01
	3	0.04	0.00	0.01
	4	0.05	0.00	0.01
	5	0.07	0.00	0.01
	6	0.08	0.01	0.01
	7	0.09	0.01	0.01
	8	0.11	0.01	0.01
	9	0.12	0.01	0.01
	10	0.13	0.02	0.01
	11	0.15	0.02	0.01
	12	0.16	0.03	0.01
	13	0.17	0.03	0.01
	14	0.19	0.03	0.01

What if now we have a continuous curve, the graph of $y = f(x)$ defined on the whole interval $[a, b]$? These estimates are exact in the case of a linear f.

We will, just as before, use a partition of the interval to split the curve into smaller pieces but then we will approximate these pieces not with horizontal segments but with *secant lines*.



Let's *define* and then compute the length of the graph of $y = f(x)$ over the interval $[a, b]$.

Then, the full length of the graph of f is approximated by the sum of all n of those, as follows:

$$\text{length } \approx L_n = \sum_{k=1}^{n} \sqrt{\Delta x_k^2 + (f(x_k) - f(x_{k-1}))^2}\,.$$

The limit of this sequence is the meaning of the length of the curve!

We would prefer, however, to connect this idea back to the idea of the Riemann integral and to the machinery that we have developed. Unfortunately, this expression doesn't look like the Riemann sum of a function! What is missing is Δx_k as a multiple in each of the terms. We will have to create it by manipulating the formula.

We can use the insight from the earlier discussion: There must be the derivative of f present. This means that we must see the *difference quotient* in the formula! Where is it? We see the difference but not the difference *quotient*. We will need to create it by manipulating the formula.

The two goals match up: We divide and multiply each term by Δx_k. Two birds with one stone:

$$\begin{aligned}
\text{Sum of lengths } &= \sum_{k=1}^{n} \sqrt{\Delta x_k^2 + (f(x_k) - f(x_{k-1}))^2}\\
&= \sum_{k=1}^{n} \sqrt{\Delta x_k^2 + (f(x_k) - f(x_{k-1}))^2} \cdot \frac{\Delta x_k}{\Delta x_k}\\
&= \sum_{k=1}^{n} \sqrt{\frac{1}{\Delta x_k^2}\left(\Delta x_k^2 + (f(x_k) - f(x_{k-1}))^2\right)} \cdot \Delta x_k \quad \text{Here is } \Delta x_k\,.\\
&= \sum_{k=1}^{n} \sqrt{1 + \left(\frac{f(x_k) - f(x_{k-1})}{\Delta x_k}\right)^2} \cdot \Delta x_k\,. \quad \text{Here is the difference quotient.}
\end{aligned}$$

But this is still not the Riemann sum. The expression that precedes Δx_k would have to be the value of some function evaluated at the *secondary nodes* of the partition. We haven't specified those yet and that's a good news because now it is our choice!

We apply, as we've done many times before, the *Mean Value Theorem*. There is some c_k in the interval $[x_{k-1}, x_k]$ such that the slope of the tangent line at that location is equal to the slope of the secant line over the interval:

$$\frac{f(x_k) - f(x_{k-1})}{\Delta x_k} = f'(c_k)\,.$$

Therefore,

$$\text{Sum of lengths } = \sum_{k=1}^{n} \sqrt{1 + (f'(c_k))^2} \cdot \Delta x_k\,.$$

Finally, this *is* the Riemann sum of a function over the partition of $[a, b]$ with the secondary nodes $c_1, ..., c_n$; the function is:

$$g(x) = \sqrt{1 + (f'(x))^2}\,.$$

Just as for the area (mass, work, etc.), the analysis above reveals the *meaning* of the new concept.

Definition 3.9.6: length of curve

The *length of the curve* given by the graph $y = f(x)$ of a differentiable function over interval $[a, b]$ is defined to be the following integral:

$$\boxed{\text{Length } = \int_a^b \sqrt{1 + (f')^2}\, dx}$$

if it exists.

Note that the function f itself is *absent* from the formula! That's understandable because only the shape (given by the derivative) and not the location matters for the length of the curve. In fact, we know from Chapter 2DC-5 that if $f' = g'$, then $f = g + C$ and its graph has the same length.

Theorem 3.9.7: Length of Curve

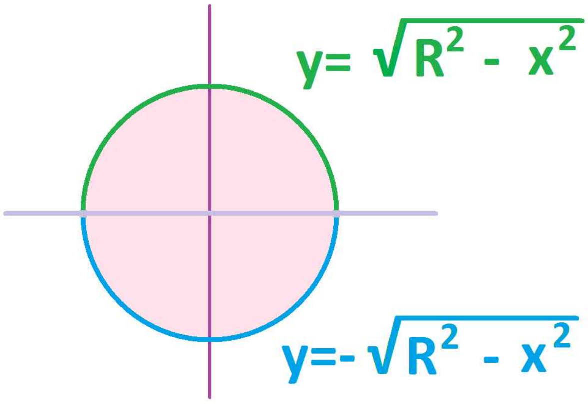
If the derivative of a function f is continuous, the length of the curve given by the graph $y = f(x)$ over $[a,b]$ is defined.

Proof.

We need the extra condition to ensure that the *Mean Value Theorem* applies and the resulting function is integrable.

Example 3.9.8: circumference of circle

It is time to prove that the circumference of a circle of radius R is $2\pi R$.



We represent, again, the upper half of the circle by the graph:

$$y = f(x) = \sqrt{R^2 - x^2}\,.$$

Then,

$$f'(x) = -\frac{x}{\sqrt{R^2 - x^2}}\,.$$

We apply the formula:

$$\begin{aligned}
\text{Half of the length} \quad &= \int_a^b \sqrt{1 + (f'(x))^2}\, dx \\
&= \int_{-R}^{R} \sqrt{1 + \left(-\frac{x}{\sqrt{R^2 - x^2}}\right)^2}\, dx \\
&= \int_{-R}^{R} \sqrt{1 + \frac{x^2}{R^2 - x^2}}\, dx \\
&= \int_{-R}^{R} \sqrt{\frac{R^2}{R^2 - x^2}}\, dx \\
&= R \cdot \int_{-R}^{R} \frac{1}{\sqrt{R^2 - x^2}}\, dx \\
&= \ldots \\
&= R \cdot \pi\,. && \text{Via trig substitution.}
\end{aligned}$$

Exercise 3.9.9

Find the length of the segment of the parabola $y = x^2$ from $(0,0)$ to $(1,1)$.

Exercise 3.9.10

Find the length of the segment of the curve $y = x^3$ from $(0,0)$ to $(1,1)$.

Exercise 3.9.11

Find the length of the segment of the curve $y = \sin x$ above the interval $[0,\pi]$.

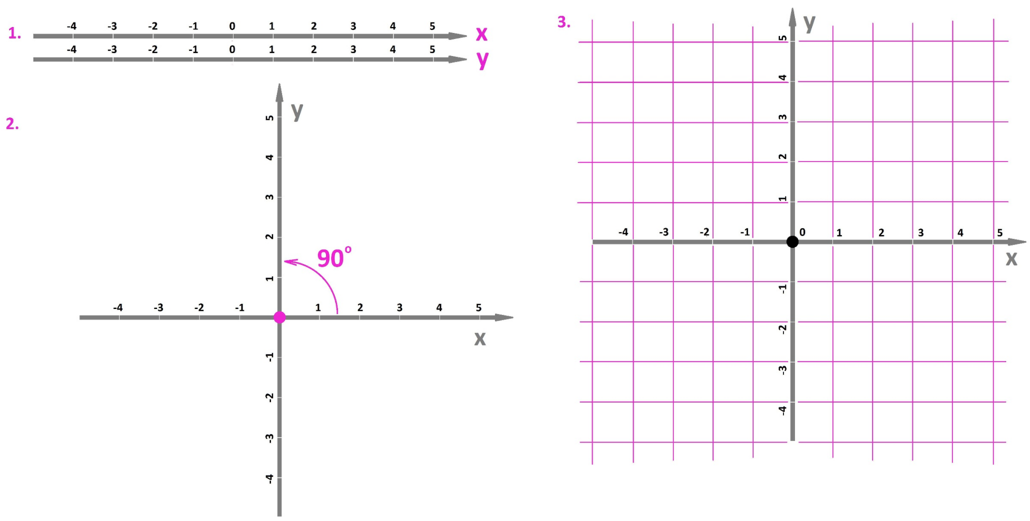
3.10. *The coordinate system for dimension three*

We pursued the idea of a coordinate system in order to transition from

- *geometry*: points, lines, triangles, circles, planes, cubes, spheres, etc., to
- *algebra*: numbers, combinations of numbers, functions, etc.

This approach allows us to solve geometric problems – such as finding the distance between two points – without measuring.

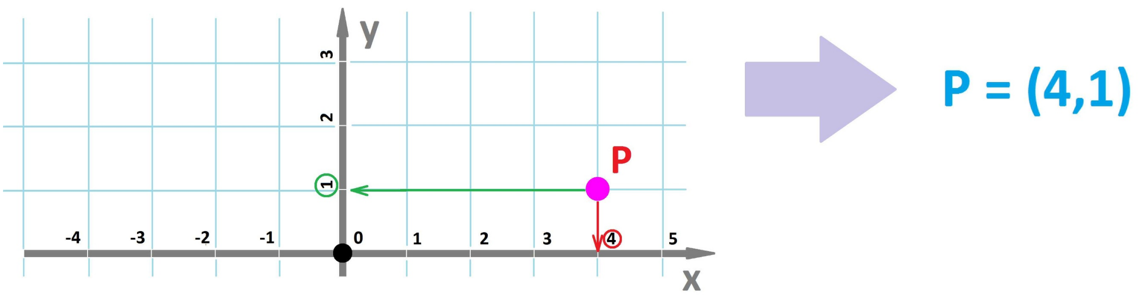
Recall how, for dimension 2, the coordinate system is built:



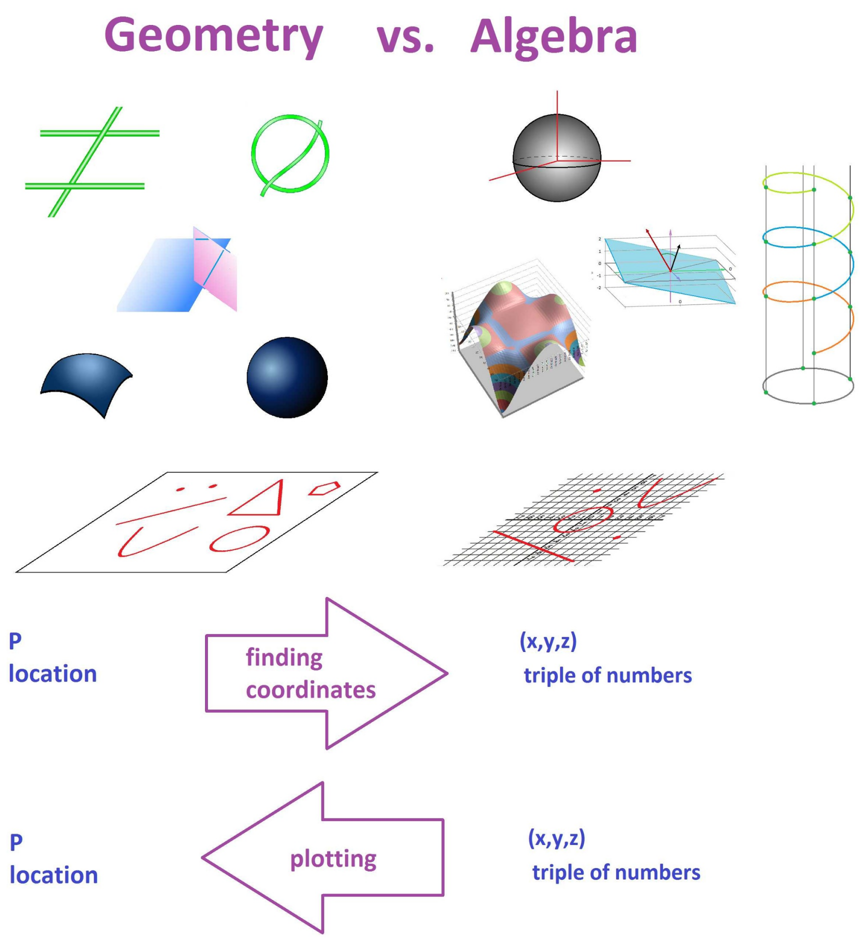
We have a correspondence:

$$\boxed{\text{location } P \longleftrightarrow \text{ pair } (x,y).}$$

This is how it works:

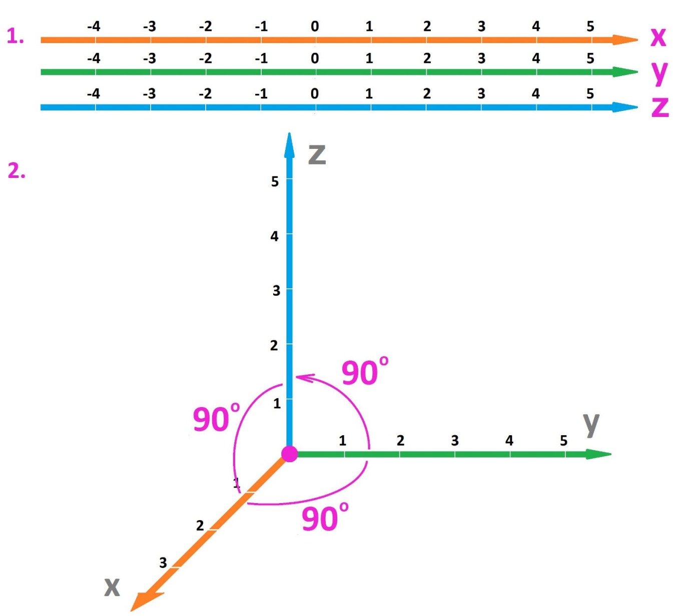


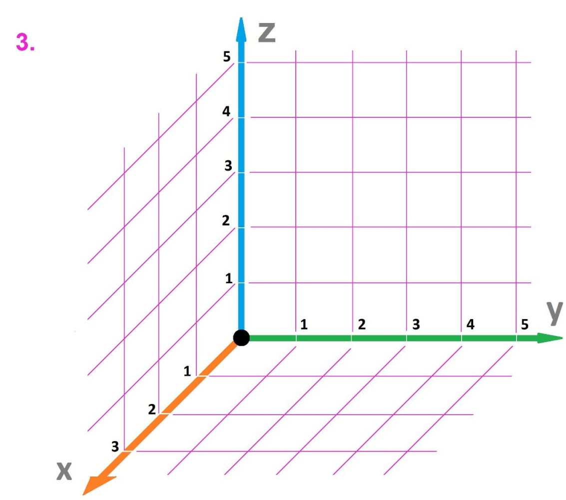
We continue with *dimension* 3. There is much more going on:



It is built in several stages:

1. Three *coordinate axes* are chosen: the x-axis, the y-axis, and the z-axis.
2. The two axes are put together at their origins so that it is a 90-degree turn from the positive direction of one axis to the positive direction of the next – from x to y to z to x.
3. Use the marks on the axis to draw a grid.





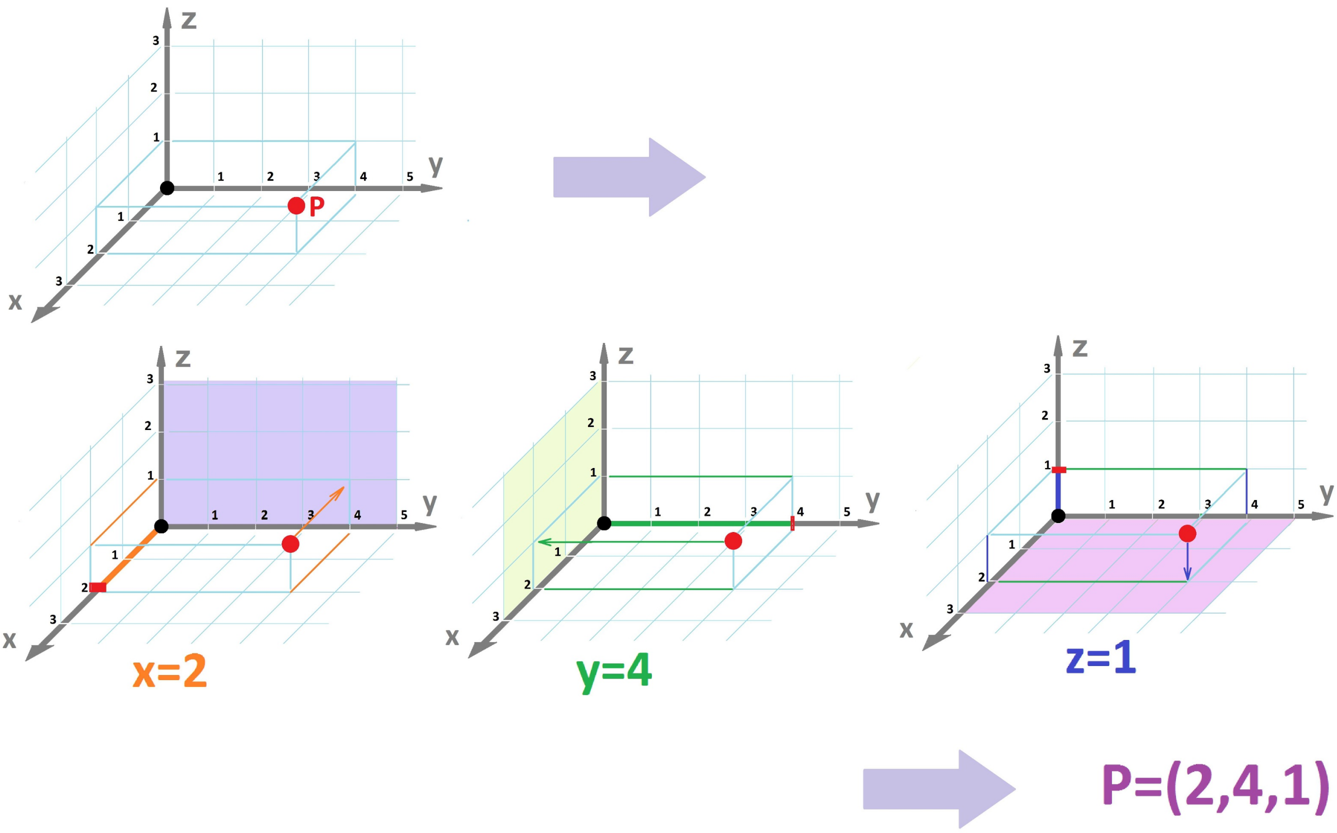
Alternatively, the system is built from three copies of the Cartesian plane: the xy-plane, the yz-plane, and the zx-plane. They are called the *coordinate planes*.

We have, as before, a correspondence between locations and their algebraic representations:

$$\boxed{\text{location } P \longleftrightarrow \text{ triple } (x,y,z)}$$

It works in *both directions.*

For example, suppose P is a *location* in this space. We then find the distances from the three planes to that location – positive in the positive direction and negative in the negative direction – and the result is the three coordinates of P, some *numbers* x, y, and z:

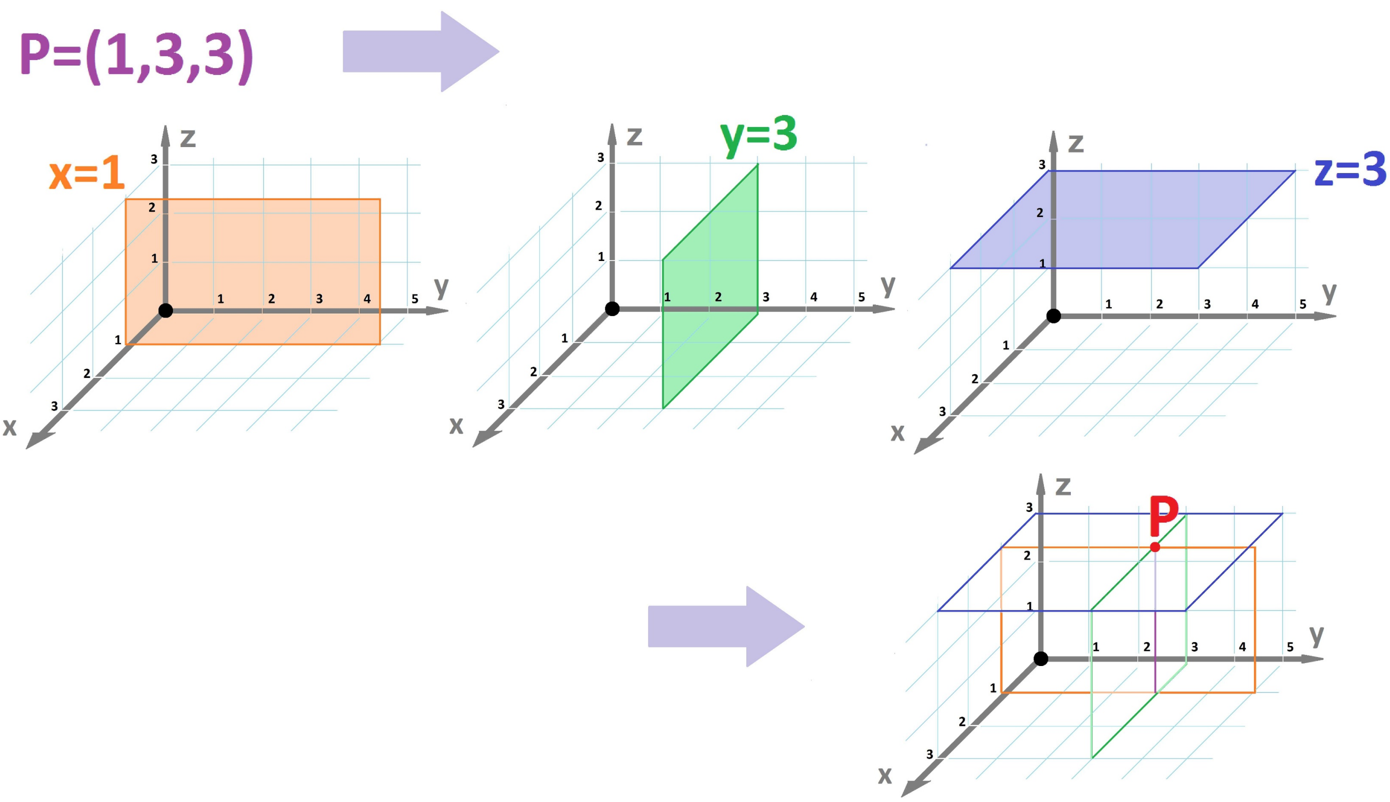


The distance from the yz-plane is measured along the x-axis, etc. We use the nearest mark to simplify the task.

Conversely, suppose x, y, z are *numbers.*

- First, we measure x as the distance from the yz-plane – positive in the positive direction and negative in the negative direction – along the x-axis and create a plane parallel to the yz-plane.
- Second, we measure y as the distance from the xz-plane along the y-axis and create a plane parallel to the xz-plane.
- Third, we measure z as the distance from the xy-plane along the z-axis and create a plane parallel to the xy-plane.

The intersection of these three planes – as if these were the two walls and the floor in a room – is a *location* $P = (x, y, z)$ in the space:

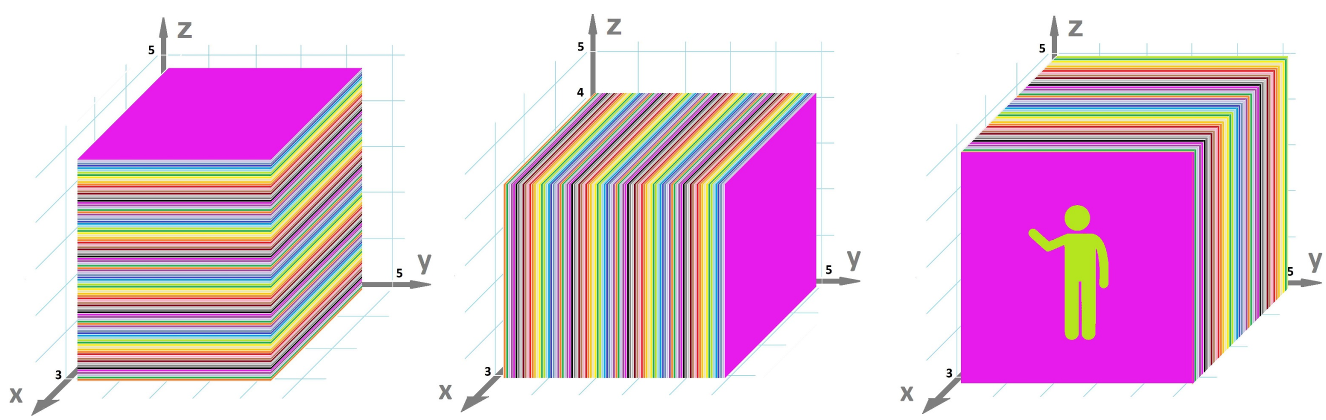


This 3-dimensional coordinate system is called *the Cartesian space*, or the 3-*space*.

Once the coordinate system is in place, it is acceptable to think of location as triples of numbers and vice versa. In fact, we can write:

$$P = (x, y, z) .$$

One can think of the 3-space as a *stack of planes*, each of which is just a copy of one of the coordinate planes:



We can use this idea to reveal the internal structure of the space.

Theorem 3.10.1: Planes Parallel to Coordinate Planes

1. *If L is a plane parallel to the xy-plane, then all points on L have the same z-coordinate. Conversely, if a collection L of points consists of all points with the same z-coordinate, L is a plane parallel to the xy-plane.*
2. *If L is a plane parallel to the yz-plane, then all points on L have the same x-coordinate. Conversely, if a collection L of points consists of all points with the same x-coordinate, L is a plane parallel to the yz-plane.*
3. *If L is a plane parallel to the zx-plane, then all points on L have the same y-coordinate. Conversely, if a collection L of points consists of all points with the same y-coordinate, L is a plane parallel to the zx-plane.*

Then, we have a compact way to represent these planes:

$$x = k, \ y = k, \quad \text{or } z = k\,,$$

for some real k.

Relations are used in the same way as before but with more variables. A relation processes a triple of numbers (x, y, z) as the input and produces an output, which is: Yes or No. If we are to plot the *graph* of a relation, this output becomes: a point or no point.

For example, let's try one of the relations above:

$$\begin{array}{ccccc} \text{input} & & \text{relation} & & \text{output} \\ (x,y,z) & \mapsto & \boxed{x = 2?} & \xrightarrow{\texttt{TRUE}} & \text{Plot } (x,y,z)\,. \\ & & \downarrow{\scriptstyle\texttt{FALSE}} & & \\ & & \text{Don't plot.} & & \end{array}$$

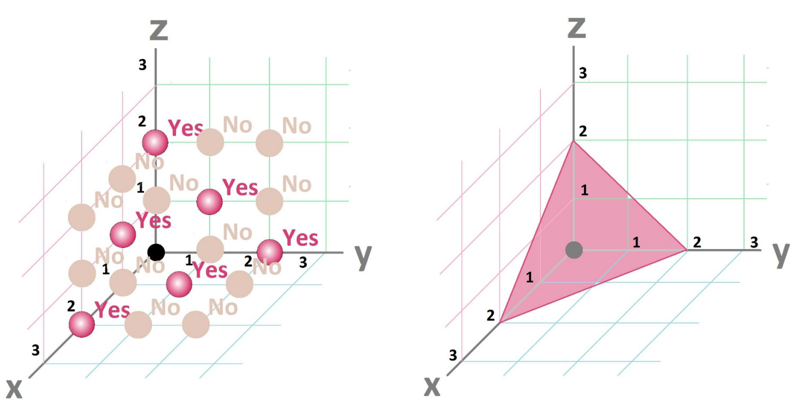
Only the points with the x-coordinate equal to 2 will be plotted.

Example 3.10.2: plane

Consider this relation:

$$\begin{array}{ccccc} \text{input} & & \text{relation} & & \text{output} \\ (x,y,z) & \mapsto & \boxed{x + y + z = 2?} & \xrightarrow{\texttt{TRUE}} & \text{Plot } (x,y,z)\,. \\ & & \downarrow{\scriptstyle\texttt{FALSE}} & & \\ & & \text{Don't plot.} & & \end{array}$$

We can do it by hand:



We can use, as before, the *set-building notation*:

$$\{(x, y, z) : \quad \text{a condition on } x, y, z\}\,.$$

For example, the graph of the above relation is a subset of $\mathbf{R}^2$ given by:

$$\{(x, y, z) : \ x + y + z = 2\}\,.$$

We accept the following without proof:

Theorem 3.10.3: Plane

Every plane through a point (h,k,l) is given by the relation:

$$A(x-h)+B(y-k)+C(z-l)=0\,.$$

In particular, we have:

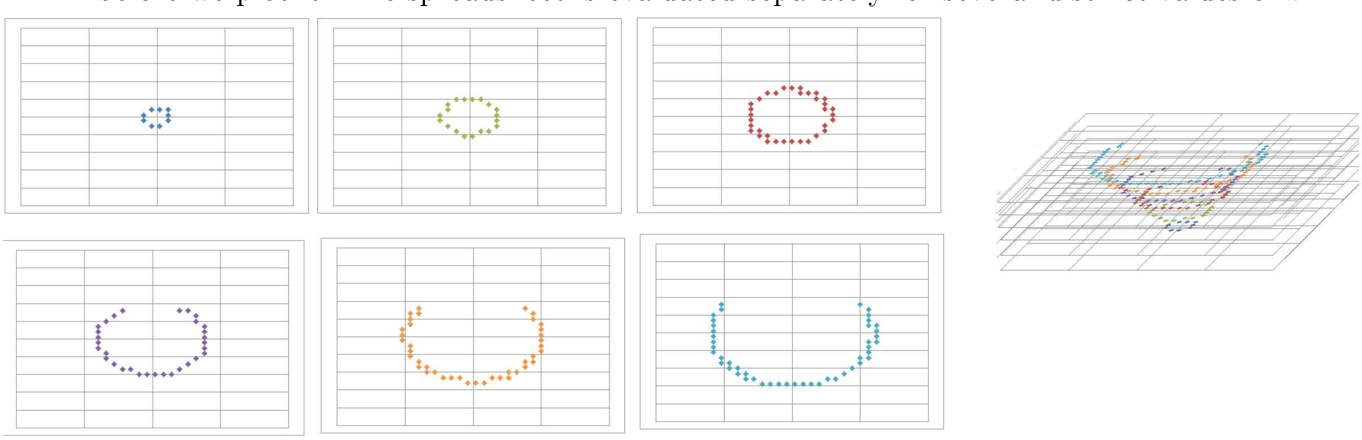
1. plane parallel to the yz-plane: $B=C=0$
2. plane parallel to the zx-plane: $A=C=0$
3. plane parallel to the xy-plane: $A=B=0$

Example 3.10.4: equal distance

How do we plot the graph of a more complex relation? Let's consider this:

input	relation		output
(x,y,z) $\mapsto$	$\boxed{x^2+y^2+z^2=1?}$	$\xrightarrow{\texttt{TRUE}}$	Plot $(x,y,z)\,.$
	$\downarrow$ `FALSE`		
	Don't plot.		

We test each of these triples (x,y,z) with the help of a spreadsheet. Just as before, instead of testing whether $x^2+y^2+z^2$ is equal to 1, we check whether it is within a small fixed number, such as 0.001, from 1 before we plot it. The spreadsheet is evaluated separately for several distinct values of z:

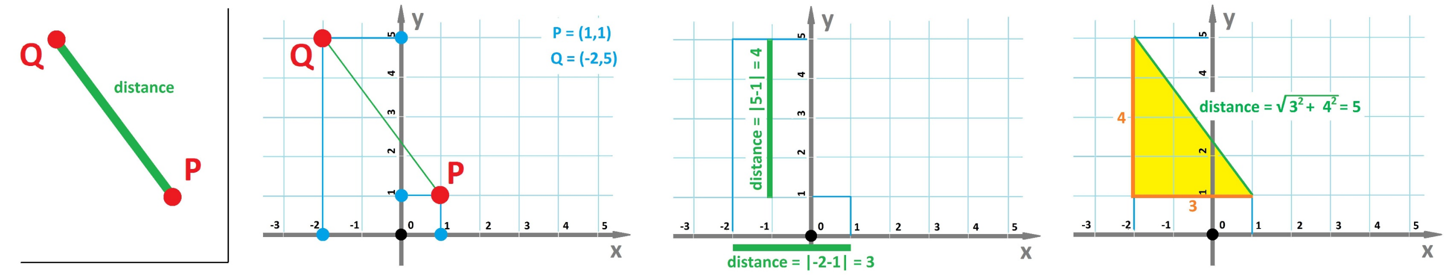


Then we put these together as sheets of paper (far right). The result looks like a surface; we will demonstrate that we have a sphere.

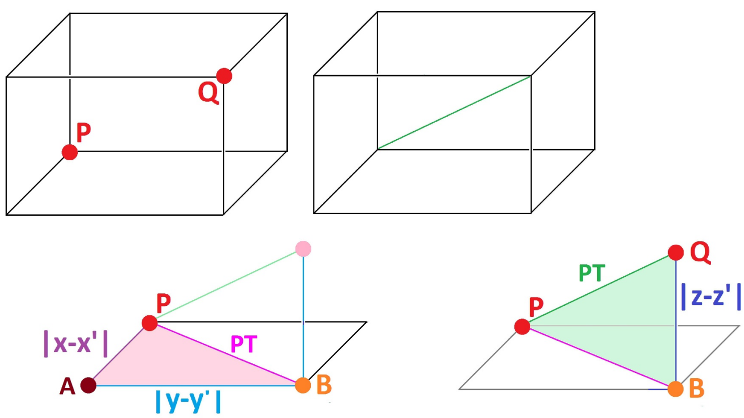
Now that everything is *pre-measured*, we can solve the geometric problems by algebraically manipulating coordinates.

The first geometric task is finding the *distance*. What is the distance between locations P and Q in terms of their coordinates (x,y,z) and (x',y',z')?

For dimension 2, we used the distance formula from the 1-dimensional case. We found the distance between two points on the plane as the length of the diagonal of the rectangle – with its sides parallel to the coordinate axes – that has these points at the opposite corners:



Similarly, we find the distance between two points in space as the length of the diagonal of the *box* – with its edges parallel to the coordinate axes and sides parallel to the coordinate planes – that has these points at the opposite corners:



Theorem 3.10.5: Distance Formula for Dimension 3

The distance between points given by their coordinates $P=(x,y,z)$ and $Q=(x',y',z')$ is the following:

$$\boxed{d(P,Q)=\sqrt{(x-x')^2+(y-y')^2+(z-z')^2}}$$

Proof.

We use the distance formula from the 1-dimensional case separately for each of the three axes, as follows. The distance

- between x and x' on the x-axis is $|x-x'|$,
- between y and y' on the y-axis is $|y-y'|$, and
- between z and z' on the z-axis is $|z-z'|$.

Then, the segment between the points $P=(x,y,z)$ and $Q=(x',y',z')$ is the diagonal of this "box". Its sides are: $|x-x'|$, $|y-y'|$, and $|z-z'|$. Our conclusion below follows from the *Pythagorean Theorem* applied twice: We first find the length of the diagonal of the opposite face of the box and then the length of the main diagonal, as follows:

$$\begin{array}{llll}
d(P,A)=|x-x'| & d(A,B)=|y-y'| \implies & d(P,B)^2 & =(x-x')^2+(y-y')^2 \\
d(P,B)^2=(x-x')^2+(y-y')^2 & d(B,Q)=|z-z'| \implies & d(P,Q)^2 & =d(P,B)^2+d(B,Q)^2 \\
 & & & =(x-x')^2+(y-y')^2+(z-z')^2
\end{array}$$

Exercise 3.10.6

Prove that in the latter case the triangle is indeed a right triangle.

A treatment of the second geometric task, *directions*, is postponed until Chapter 4HD-1.

This is our conclusion about the relation considered above:

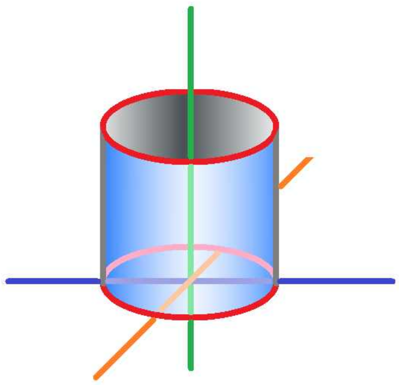
Theorem 3.10.7: Sphere

The sphere of radius $R > 0$ centered at a point (h, k, l), which is the collection of all points R units away from (h, k, l), is given by the relation:

$$(x-h)^2 + (y-k)^2 + (z-l)^2 = R^2 .$$

Proof.

It follows from the *Distance Formula* for dimension 3.



Theorem 3.10.8: Cylinder

The cylinder of radius $R > 0$ centered around the z-axis, which is the collection of all points R units away from the axis measured horizontally, is given by the relation:

$$x^2 + y^2 = R^2 .$$

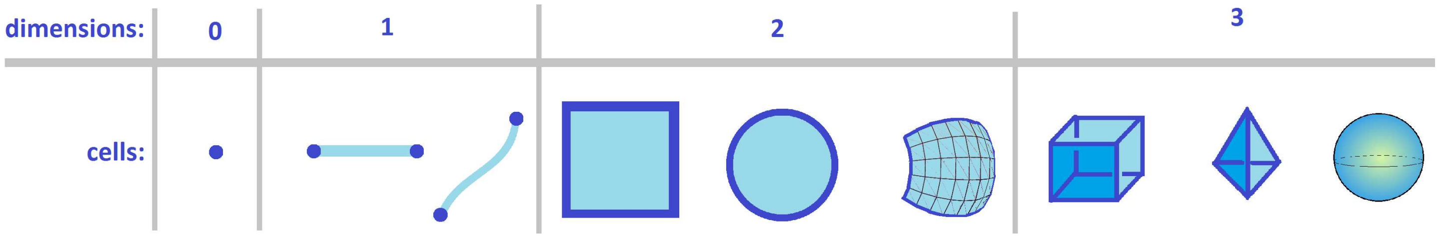
Proof.

It follows from the *Distance Formula for Dimension* 2.

Representations of sets of all points 1 unit away from the origin are presented below, for the *dimensions of the space* $n = 1, 2, 3$:

dimension:	1	2	3
distance $=1$	$\lvert x\rvert = 1$	$x^2 + y^2 = 1$	$x^2 + y^2 + z^2 = 1$
set:	two points	circle	sphere
its "size":	N/A	length	area
distance ≤ 1	$\lvert x\rvert \le 1$	$x^2 + y^2 \le 1$	$x^2 + y^2 + z^2 \le 1$
set:	interval	disk	ball
its "size":	length	area	volume

The latter list is a list of the building blocks called "cells". They are presented below, for the *dimensions of the cells* $m = 0, 1, 2, 3$:



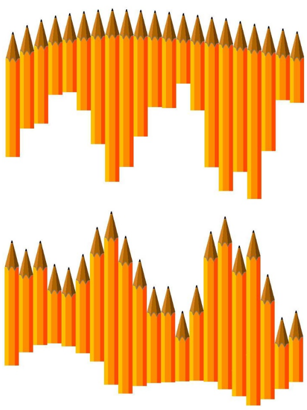
The transition from relations of three variables to *functions* of two variables will be discussed in Chapter 4.

We see how much harder it is to visualize things in the 3-dimensional space. That's why we will need a further development of the algebraic treatment of these geometric ideas in Volume 4 (Chapter 4PC-2).

3.11. Volumes of solids via their cross-sections

We have come to understand *areas in terms of lengths*.

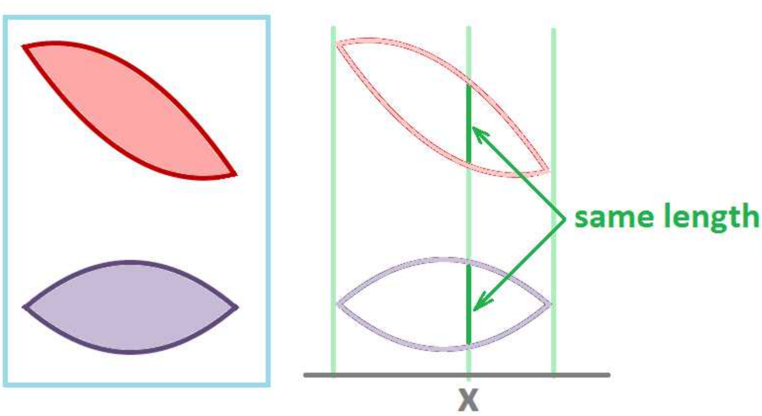
Indeed, if we rearrange these pencils by moving each up or down, they will still cover the same area:



This fact is meant to illustrate the following situation. Suppose we have four functions f, g, F, G that have nothing to do with each other except the distance between the graphs – on the xy-plane – is the same:

$$f(x) - g(x) = F(x) - G(x) ,$$

for all x in $[a, b]$. Let's compare: the area between f and g vs. the area between F and G. Each pair of corresponding rectangles in the approximations of the two areas over some partition have the *same height* (same pencil).



That is why the Riemann sums over a partition of $[a, b]$ that approximate the areas between either pair of graphs are the same:

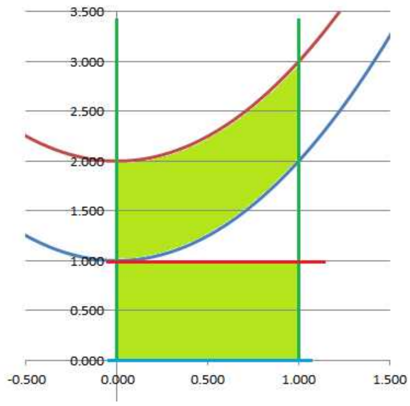
$$\Sigma (f - g) \cdot \Delta x = \Sigma (F - G) \cdot \Delta x ,$$

over any augmented partition P. Therefore, the integrals – the areas between the graphs – are equal too:

$$\int_a^b (f-g)\,dx = \int_a^b (F-G)\,dx\,.$$

Example 3.11.1: equal areas

The area between the graphs of $y = x^2+1$ and $y = x^2+2$ is the same as that of the square below:



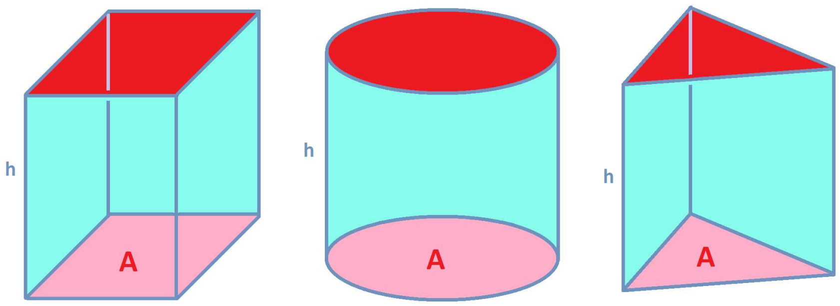
Conclusion:

▶ The vertical *cross-section of the plane region* bounded by $y = f(x)$ and $y = g(x)$ is the vertical segment $[g(x), f(x)]$ for each x, and only the lengths of these segments, $f(x) - g(x)$, affect the area of the region.

Let's now go up in dimension and examine the *cross-sections of a solid* and find out what they tell us about the volume of this solid.

But first, what is *volume*? The question will be addressed in full only in Volume 4 (Chapter 4HD-5). For now, we will rely on a simplifying assumption.

We do understand the meaning of the volume of such a simple solid as a *box*. It is $V = w \cdot d \cdot h$, where w is the width, d the depth, and h the height. We also "know" the volume of a cylinder, $V = \pi R^2 h$, where R is its radius and h is the height.

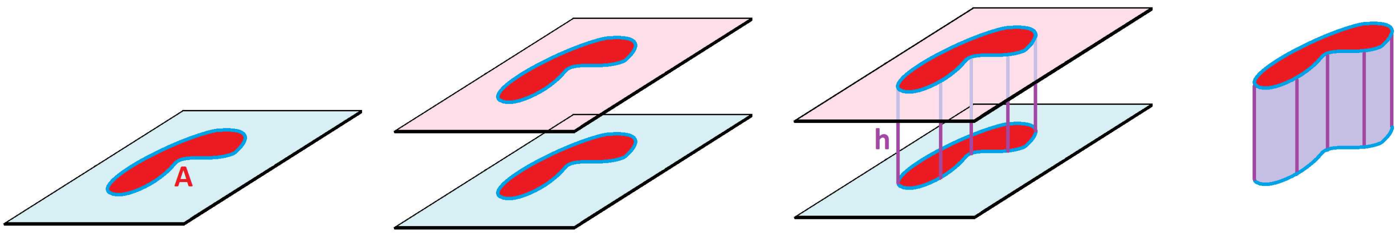


We can gain insight from this:

$$\boxed{\text{Volume} \;=\; \text{area of the base} \;\cdot\; \text{height}}$$

Indeed, the area of the base is, respectively, $A = wd$ and $A = \pi R^2$. The same is true for the prism.

What do they all have in common? The base is a region on the plane and we might know its area A (an integral). This region is lifted off the plane to the height h. Between these two plane regions lies a cylinder-like solid:



We will *assume* that its volume is:

$$V = A \cdot h\,.$$

Just as we have been using rectangles to approximate the slices of plane regions, these "shells" will approximate slices of solids.

Suppose that, instead of a stack of pencils, we have a stack of *coins*. If we rearrange these coins by moving them side to side, the total volume will remain the same:

We realize that we should try to understand *volumes in terms of areas*. It is called the *Cavalieri principle*:

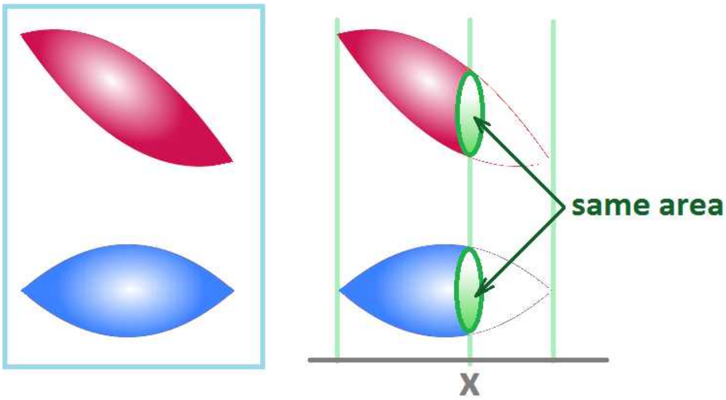
> If the vertical cross-sections of two solids in the space have equal areas, then their volumes are also equal.

To confirm that this principle makes sense, we can match it with what we know about the idea of area in terms of lengths:

$$\text{If the vertical cross-sections of two } --\left\langle \begin{array}{c} \text{regions in the plane} \\ \text{solids in the space} \end{array} \right\rangle--$$

$$\text{have equal } --\left\langle \begin{array}{c} \text{lengths} \\ \text{areas} \end{array} \right\rangle--\text{, then their } --\left\langle \begin{array}{c} \text{areas} \\ \text{volumes} \end{array} \right\rangle-- \text{ are also equal.}$$

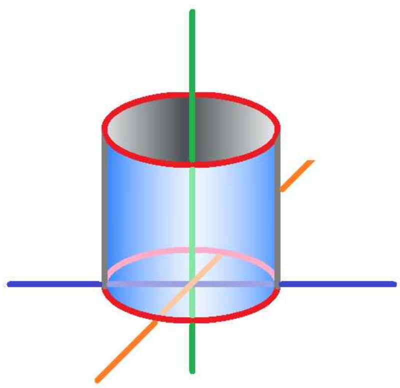
Suppose our solid S is located in the Cartesian 3-space. Its *cross-sections* are the intersections of S with the various planes, especially the ones parallel to the coordinate planes. We choose those parallel to the yz-plane and, therefore, perpendicular to the x-axis. Thus, we will consider all vertical cross-section of this solid corresponding to all values x as the intersections of S with the vertical planes through the point x on the x-axis. Each of them is a plane region and, according to the Cavalieri principle, only its area affects the volume of the region:



We denote this area of the cross-section at x by $A(x)$. It is simply a function of x.

Example 3.11.2: cylinder

What is the volume of a cylinder of radius R and height h?



It is located in our 3-space, but all we need to know is its dimensions. We have

$$A(x) = \pi R^2 .$$

By the Cavalieri principle, the volume of this cylinder is the same as the volume of a box *the* cross-section of which is a square with area πR^2 and the same height:

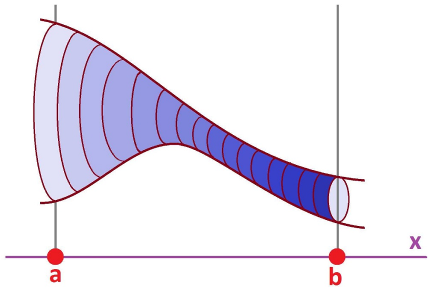
$$\text{Volume } = \pi R^2 \cdot h .$$

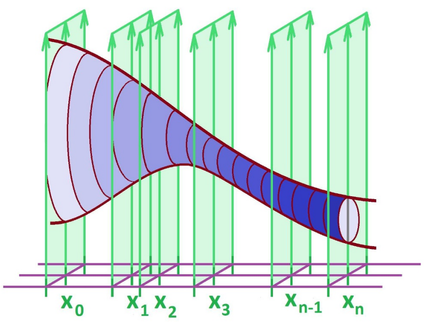
Let's confirm the idea of the Cavalieri principle via Riemann sums.

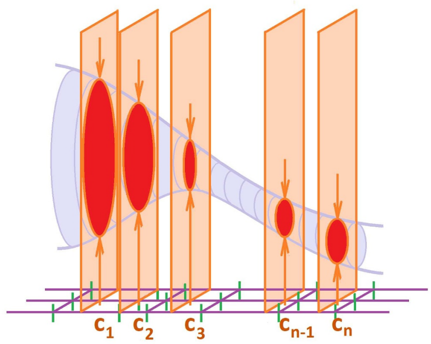
We place the x-axis somehow along the solid. Suppose the solid S lies entirely between some vertical planes $x = a$ and $x = b$. We continue with an augmented partition P of the interval $[a, b]$:

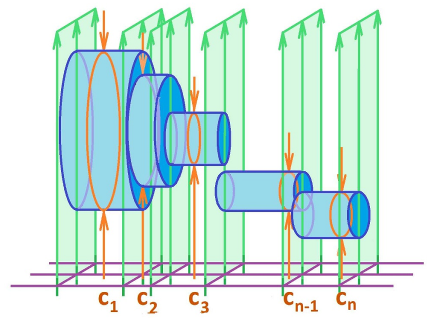
$$a = x_0 \le c_1 \le x_1 \le ... \le x_n = b$$

The vertical planes $x = x_i$ cut the solid into n slices. The ith slice is approximated by the following. The cross-section of S created by the vertical plane $x = c_i$ is a plane region; its area is $A(c_i)$:









We construct a new solid from this plane region by giving it a thickness equal to $\Delta x_i = x_i - x_{i-1}$. Then its volume is $A(c_i) \cdot \Delta x_i$. Then we have:

$$\text{Total volume } = \sum_{i=1}^{n} A(c_i) \cdot \Delta x_i \,.$$

This values is then *recognized* as the Riemann sum of $y = A(x)$ over $[a, b]$: $\Sigma\, A \cdot \Delta x$. Their limit is the Riemann integral.

Definition 3.11.3: volume of solid

The *volume of a solid* is defined to be the integral

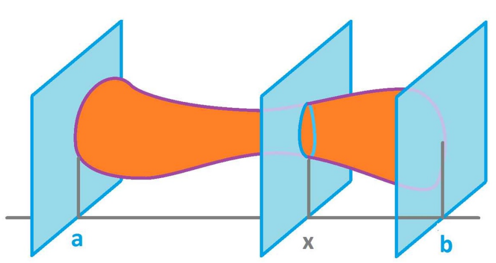
$$\int_a^b A(x)\, dx \,,$$

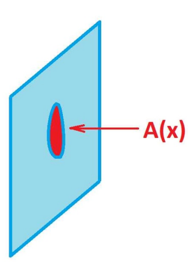
if it exists, where $A(c)$ is the area (if it exists) of the intersection of the solid and the plane $x = c$.

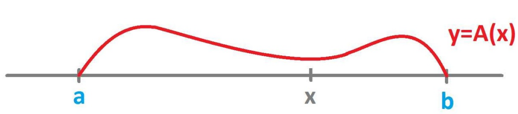
Warning!

The area $A(x)$ itself, for each x, is understood, and may have to be computed, as a Riemann integral.

Thus, the volume is the integral of the area:

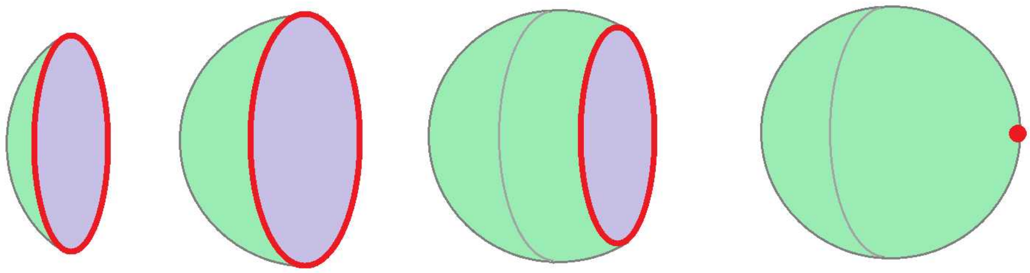




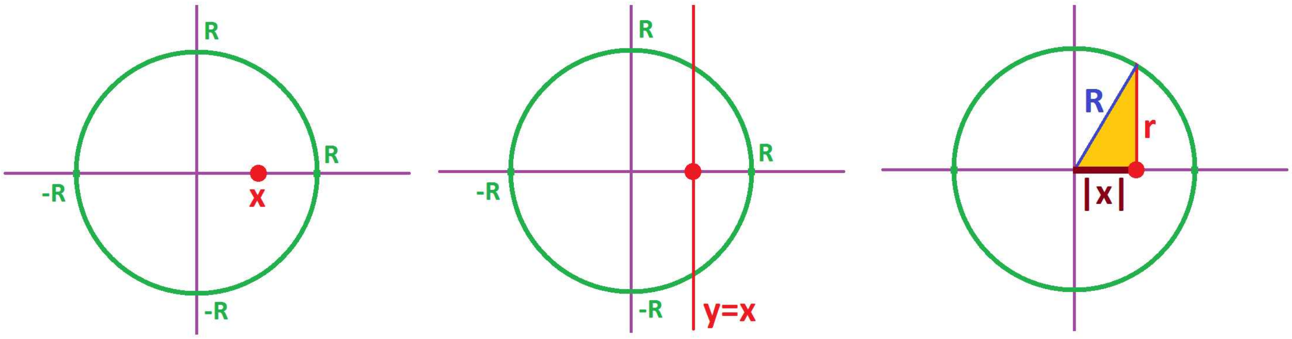


Example 3.11.4: sphere

The cross-sections of the sphere are circles:



More precisely, the cross-sections of the ball are disks and it is the areas of these disks that we need to find. Suppose the radius of this circle at x is r. What is it? Let's take a side view:



Then

$$x^2 + r^2 = R^2 .$$

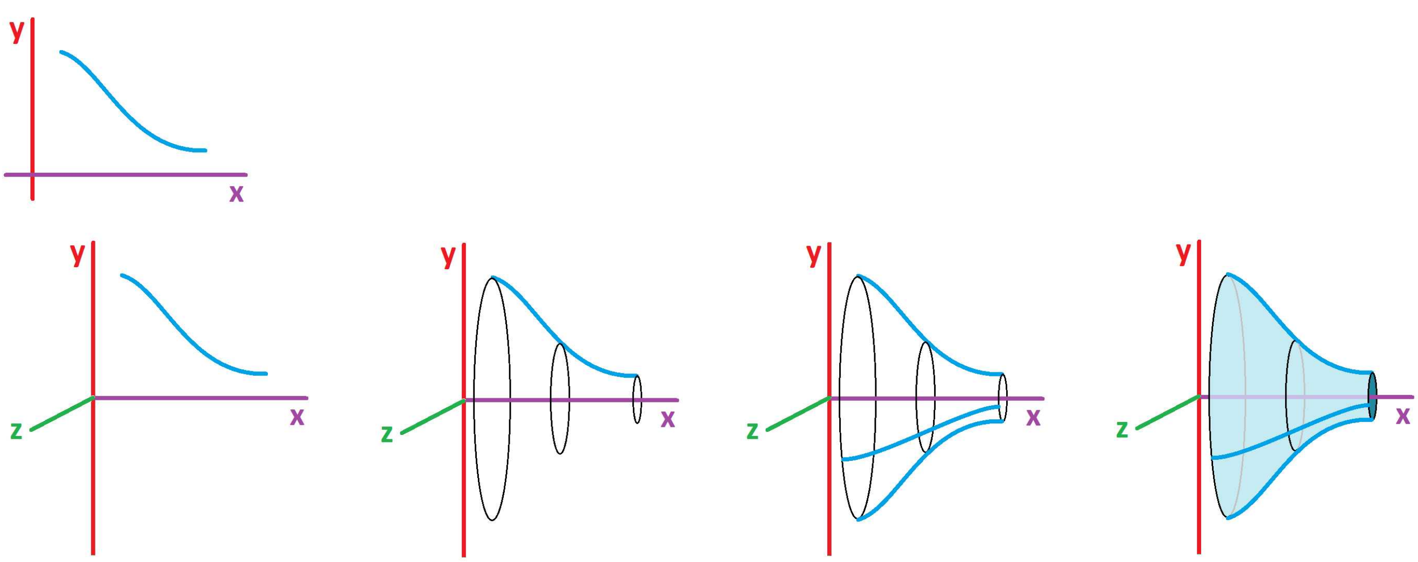
Then the area of this circle is:

$$A(x) = \pi \left(\sqrt{R^2 - x^2}\right)^2 = \pi(R^2 - x^2) .$$

Therefore,

$$\text{Volume } = \int_{-R}^{R} A(x)\, dx = \pi \int_{-R}^{R} \left(R^2 - x^2\right)\, dx = \pi \left(R^2 x - \frac{1}{3}x^3\right)\Bigg|_{-R}^{R} = \frac{4}{3}\pi R^3 .$$

We have done all preliminary work to answer the question: What is the volume when the cross-sections are circles that change from slice to slice?



Definition 3.11.5: solid of revolution

Suppose $y = f(x)$ satisfies $f(x) \geq 0$ for all x in $[a, b]$. Then, the *solid of revolution of f about the x-axis* is the following set in the xyz-space:

$$\{(x, y, z) : \sqrt{y^2 + z^2} \leq f(x)\}.$$

Theorem 3.11.6: Volume of Solid of Revolution

Suppose $y = f(x)$ satisfies $f(x) \geq 0$ for all x in $[a, b]$. Then, the volume of the solid of revolution of f about the x-axis is:

$$V = \int_a^b \pi f(x)^2 \, dx.$$

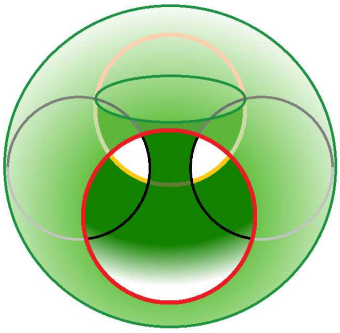
Exercise 3.11.7

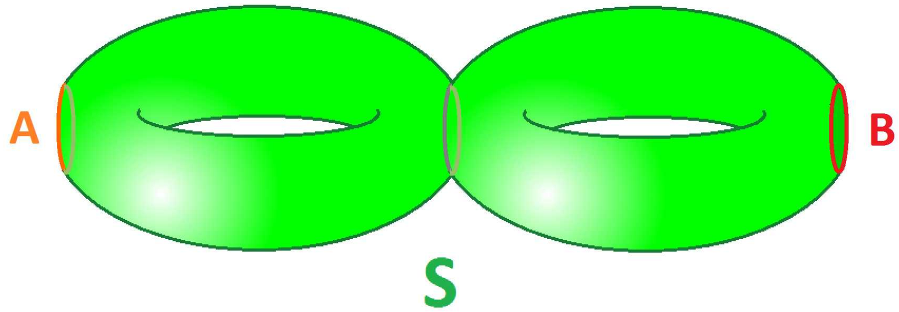
Prove the theorem.

Warning!

Even when the cross-sections are circles, they may change from slice to slice in ways that are so complex that we may have to turn to numerical integration.

In general, cross-sections can have any *geometry or topology*:



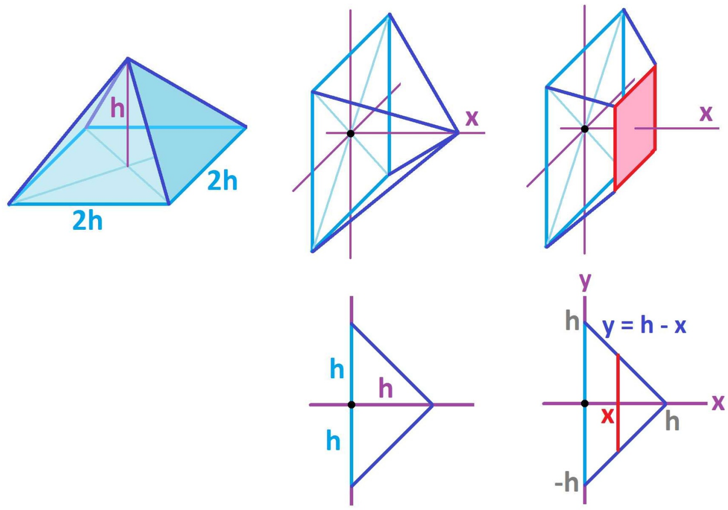


Exercise 3.11.8

Describe the cross-sections of these surfaces.

Example 3.11.9: pyramid

Let's find the volume of the right pyramid (i.e., one with its height perpendicular to its base) that has a square base with side $2h$ and height h. Its cross-sections parallel to the base are *squares*:



The side of the square located x units from the base is $2(h-x)$; therefore,

$$\text{Volume } = \int_0^h A(x)\,dx = \int_0^h 2(h-x)^2\,dx = -\frac{2}{3}(h-x)^3\Big|_0^h = \frac{2}{3}h^3\,.$$

Exercise 3.11.10

Modify the above example to find the volume of the right pyramid with square base with side Q and height h.

Exercise 3.11.11

Find the volume of the right cone with a circular base of radius R and height h.

We defined in this section the volume of a solid via its cross-sections. The definition relies on the cylindrical slices to approximate the solid. These complex objects are to be replaced with true elementary building blocks of solids – bricks and boxes – to follow the idea of the Riemann sum. The general definition is presented in Volume 4 (Chapter 4HD-5).

Exercise 3.11.12

(a) Prove that the work needed to fill – from the bottom – a tank located between the planes $x=0$ and $x=h$ (the x-axis is vertical) and with the area of its horizontal cross-section at height x equal to $A(x)$ is $\int_0^h A(x)x\,dx$. (b) Show that this work is equal to the work needed to move this mass from height 0 to the height of the center of mass of the tank.

Exercise 3.11.13

Set up but do not evaluate the Riemann sums and the integral for the volume of a box $W\times D\times H$ in terms of its cross-sections.

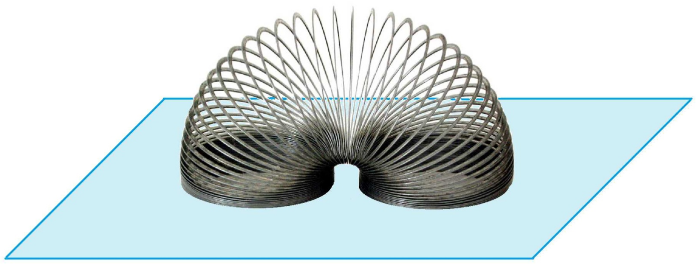
3.12. Volumes of solids of revolution

In this section, we will consider more complex surfaces of revolution.

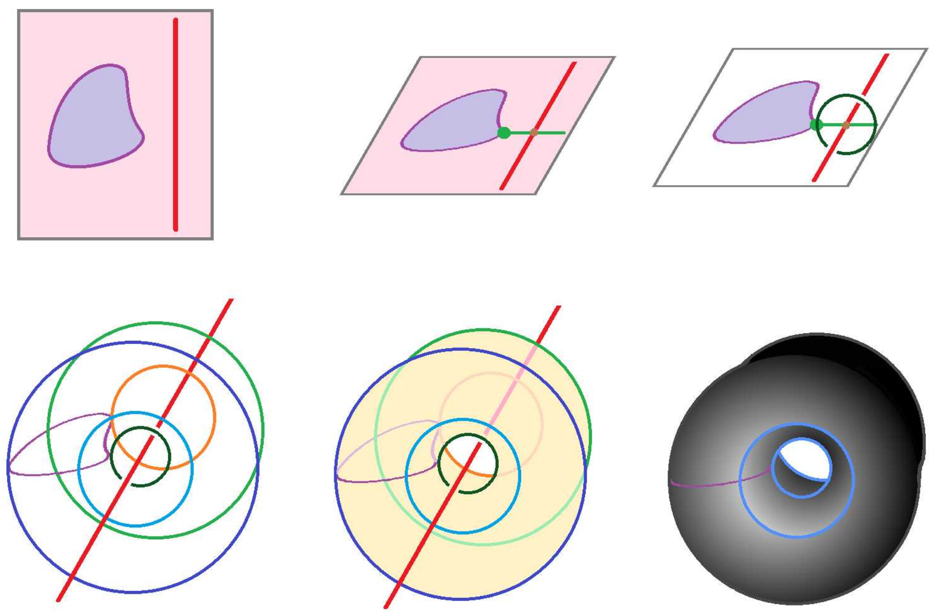
Consider an object that is rotated as it hardens:

The same effect is produced by using a cutting tool on a hard object as it is being rotated.

Let's rotate a curve. If this curve is a circle, the result of the rotation is similar to a slinky:



Mathematically, we have a curve and a line on the xy-plane, we add the z-axis, then we rotate the curve around the line in the resulting 3-space, one point at a time.

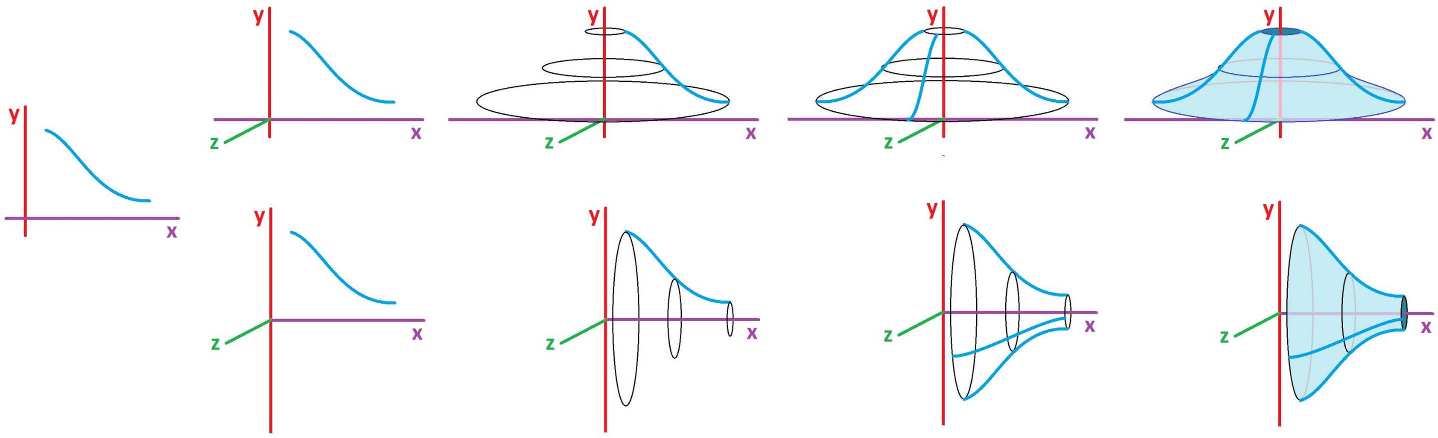


Each of the points on the curve produces a circle. Together these circles form a surface. This surface bounds a solid. What is the volume of this solid?

Suppose this curve is simply the graph of a function

$$y = f(x) \geq 0, \ a \leq x \leq b\,,$$

and suppose the line is the x-axis or the y-axis:



As the choice of the x-axis is easily addressed by the Cavalieri principle, we choose the y-axis.

Let's be clear what we are talking about. The surface created by a rotated curve has no volume; the solid it – partially – bounds does. For the case of a decreasing f, this solid contains every point (x, y, z) that satisfies:

- Its distance (measured horizontally) from the y-axis is between a and x units.
- Its distance (measured vertically) from the xz-plane is between $f(b)$ and $f(x)$ units.

Exercise 3.12.1

Describe the solid for the case of an increasing f.

The analysis of the idea of volume following the Cavalieri principle is based on cutting the solid into *disks*. Of course, it can be used for either case. Instead, we start from scratch and pursue the idea of cutting the solid into *washers* (rings).

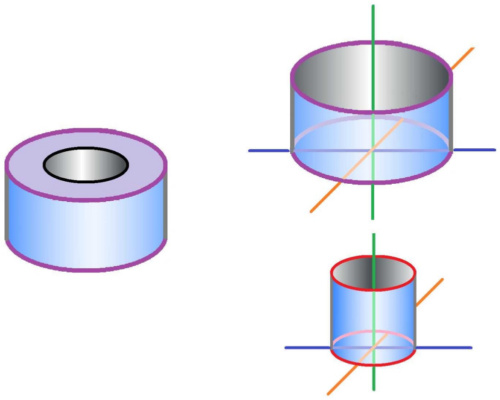
We will use, however, the following fact previously derived from the Cavalieri principle:

Theorem 3.12.2: Volume of Washer

The volume of a washer with the inner radius r, the outer radius R, and thickness h is the difference of the volumes of the two cylinders:

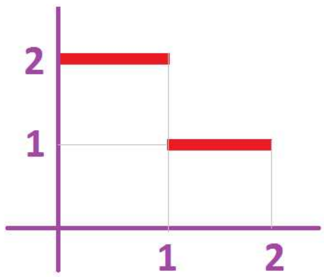
$$\text{Volume } = \pi R^2 h - \pi r^2 h = \pi h(R^2 - r^2)\,.$$

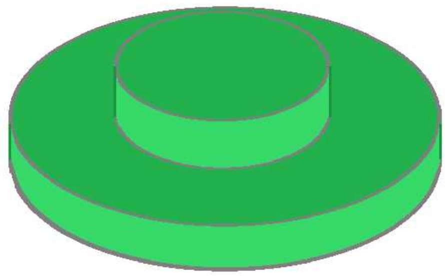
We just subtract the volumes of these two cylinders:



Example 3.12.3: pedestal

Suppose the object is simply the combination of a disk of radius 1 and the washer around it of thickness 1:



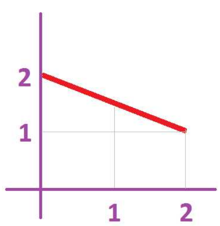


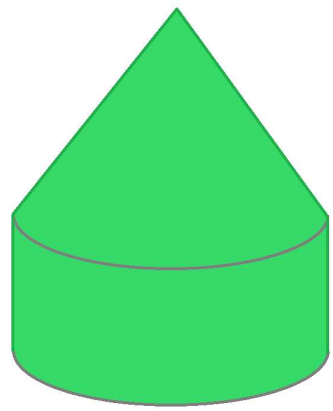
Then, the volume is simply the sum of the volume of the disk and the volume of the washer:

$$\begin{aligned} \text{Volume} \quad &= 2 \cdot \text{area of the disk} && +1 \cdot \text{area of the washer} \\ &= 2 \cdot \pi \cdot 1^2 && +1 \cdot (\pi \cdot 2^2 - \pi \cdot 1^2) \,. \end{aligned}$$

Example 3.12.4: yurt

Suppose the thickness is changing linearly from 1 to 2:





What is the volume of this object? Even though we know the answer from the Cavalieri principle, we will have to start with approximations again...

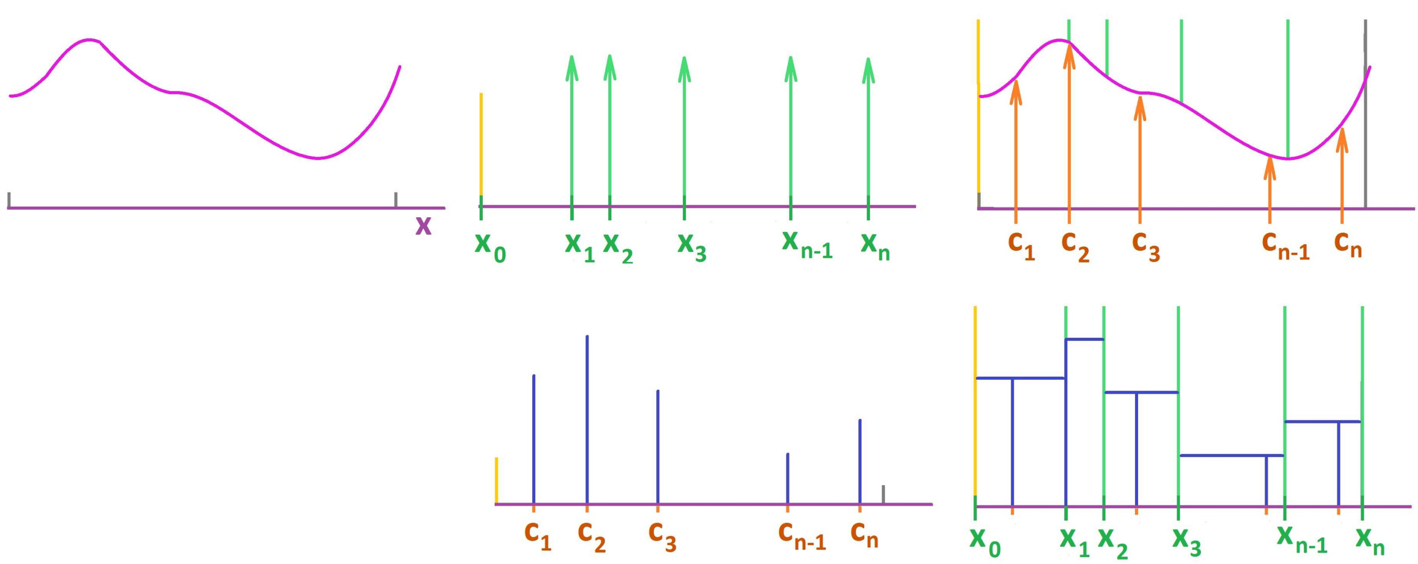
We have an augmented partition P of the radius:

$$a = x_0 \le c_1 \le x_1 \le ... \le c_n \le x_n = b$$

These are the lengths of the segments:

$$\Delta x_i = x_i - x_{i-1} \,.$$

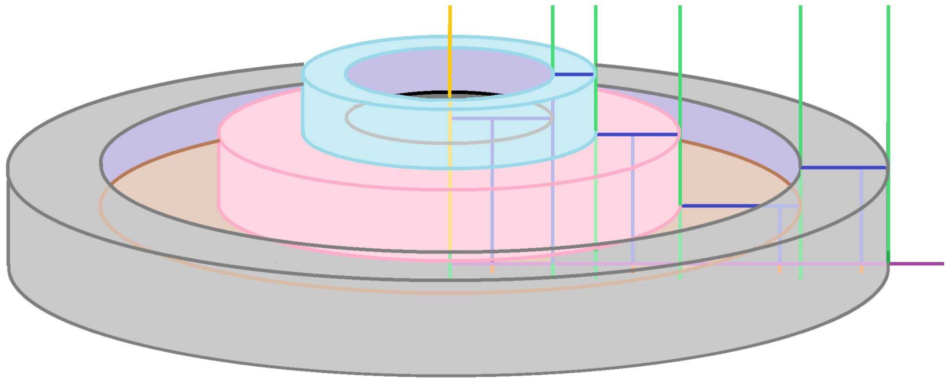
Here, we cut the solid into thin washers by the cylinders starting at points $x = x_i$ on the x-axis and then sample its height at the points c_i:



Then the height of each washer is $f(c_i)$, and we have:

$$\text{Mass of } i\text{th washer } = \text{radius} \cdot \text{area} = f(c_i) \cdot \left(\pi x_i^2 - \pi x_{i-1}^2\right) \,,$$

since the inside radius of the washer is x_{i-1} and the outside is x_i.



Then, we have:

$$\boxed{\text{Total volume } = \sum_{i=1}^{n} f(c_i) \cdot \pi \left(x_i^2 - x_{i-1}^2\right)}$$

We can use this formula for computations.

What if the solid isn't actually made of washers and its thickness varies continuously?

Then the volume of each washer – when thin enough – is *approximated* by the volume of such a washer with the constant height $f(c_i)$:

$$\text{mass of } i\text{th washer } \approx \text{radius} \cdot \text{area} = f(c_i) \cdot \left(\pi x_i^2 - \pi x_{i-1}^2\right) \,.$$

Then, we have:

$$\text{Total volume } \approx \sum_{i=1}^{n} f(c_i) \cdot \pi \left(x_i^2 - x_{i-1}^2\right) \,.$$

This is the volume of the washers built on top of the augmented partition. This time, just as on several occasions before, we *do not recognize* this expression as the Riemann sum of any function over this partition, which is supposed to be:

$$\sum_{i=1}^{n} g(c_i) \cdot \Delta x_i$$

for some function g.

Factoring takes us one step closer to the Riemann sum of a function:

$$\text{total volume } \approx \sum_{i=1}^{n} \pi f(c_i)(x_i + x_{i-1}) \cdot \Delta x_i \,.$$

We just need to do something about the term $(x_i + x_{i-1})$...

We back up a bit; we haven't chosen secondary nodes! Let's assume that function f is integrable. Then the choice of secondary nodes is ours. Let's choose the mid-points:

$$c_i = \frac{1}{2}(x_i + x_{i-1}) \,.$$

Then,

$$\text{Total volume } \approx 2\pi \sum_{i=1}^{n} f(c_i) c_i \cdot \Delta x_i \,.$$

This time, we *do* recognize this expression as the Riemann sum of a simple function. Then, we define the *volume* of the solid as the limit of these Riemann sums:

$$2\pi\Sigma x f(x)\,\Delta x\,.$$

It is important to confirm that this new definition of volume matches the old one.

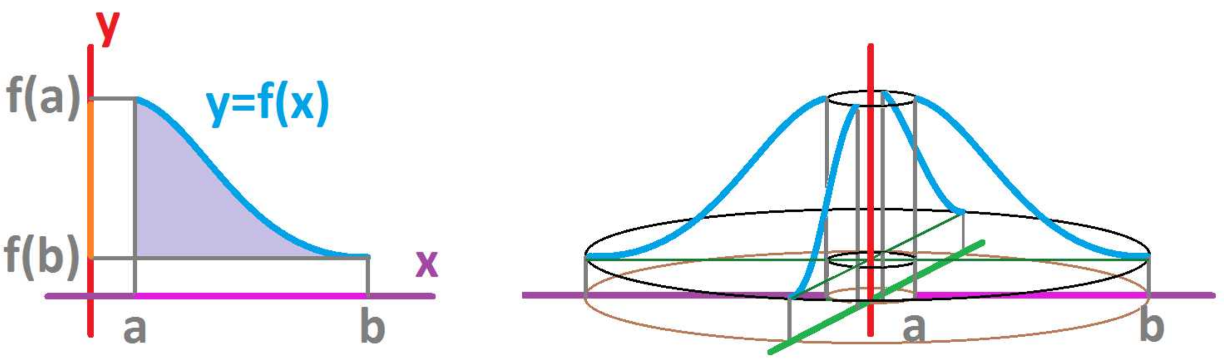
Theorem 3.12.5: Volume of Solid of Revolution

Given an integrable function f on segment $[a,b]$, the above integral is equal to the volume of the solid of revolution obtained by rotating of the graph of f:

$$\text{Volume } = 2\pi\int_a^b xf(x)\,dx\,.$$

Proof.

For simplicity, we assume that f is decreasing. We start with the definition of volume of the Cavalieri principle. The cross-sections of the solid along the y-axis are circles; specifically, the intersection of the surface with the plane $y=q$ is a circle of radius $f^{-1}(q)$.



Let's consider the whole solid swept by this curve. We know this:

$$\text{Volume } = \pi\int_{f^{-1}(b)}^{f^{-1}(a)} \left(f^{-1}(y)\right)^2\,dy\,.$$

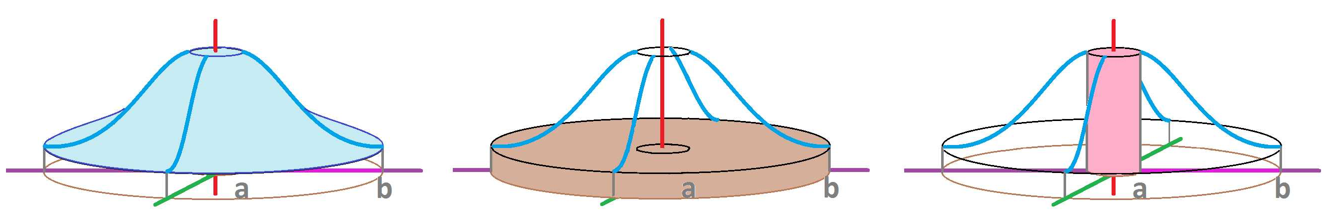
We apply *Integration by Substitution* with $x=f^{-1}(y)$. Then:

$$\text{Volume } = \pi\int_b^a x^2 f'(x)\,dx\,.$$

We apply *Integration by Parts* with $u=x^2,\ dv=f'\,dx$. Then:

$$\begin{aligned}\text{Volume } &= \pi\left(x^2f(x)\Big|_b^a - \int_b^a 2xf(x)\,dx\right)\\ &= \pi a^2 f(a) - \pi b^2 f(b) + 2\pi\int_a^b xf(x)\,dx\,.\end{aligned}$$

The extra terms come from the disk at the bottom and the cylinder in the middle to be removed:



Exercise 3.12.6

Modify the proof of the theorem for the case of an increasing f.

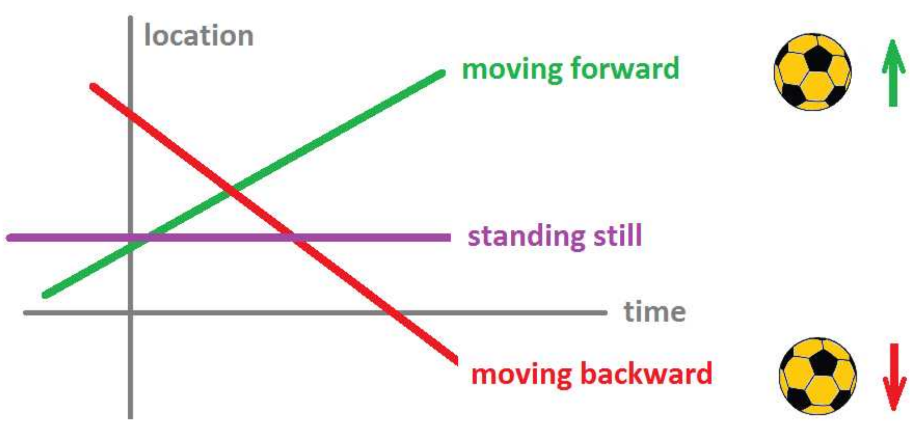
Chapter 4: Several variables

Contents

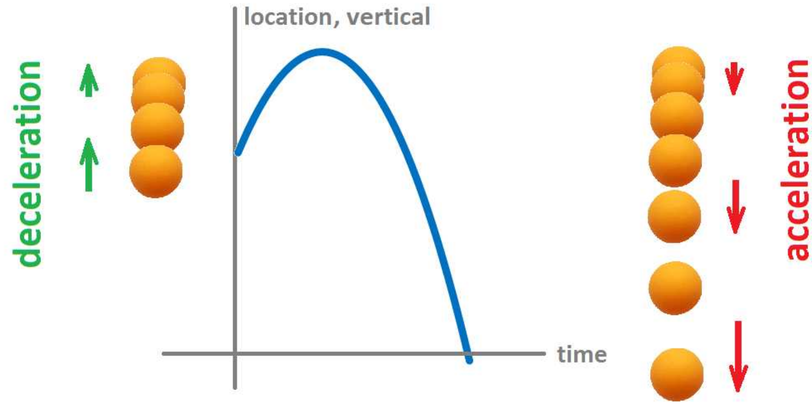
4.1. A ball is thrown...

Let's review what we learned in Chapter 2DC-3 about motion of a ball (or a cannonball).

When a ball is thrown in the air under an angle, it moves in both vertical and horizontal directions, simultaneously and independently. The dynamics is very different. In the *horizontal* direction, as there is no force changing the velocity, the latter remains constant:



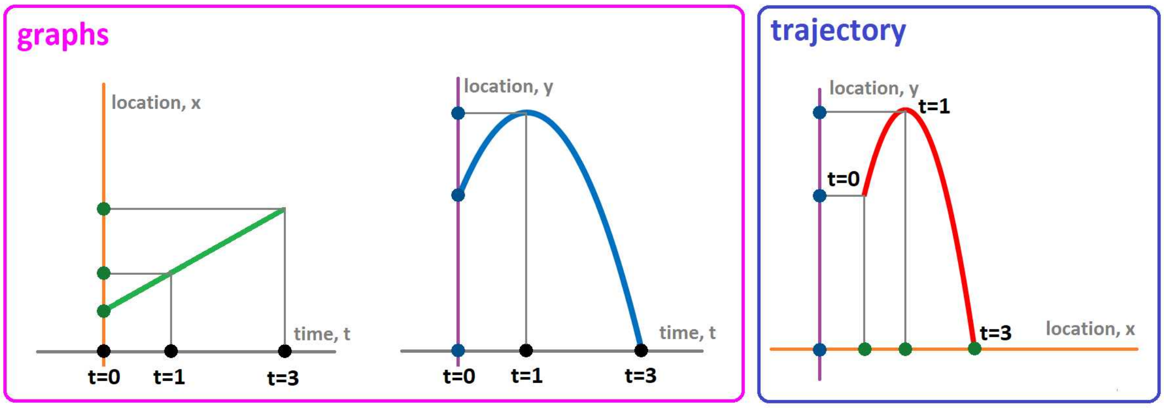
Meanwhile, the *vertical* velocity is constantly changed by the gravity. The dependence of the height on the time is quadratic:



We have *three* variables:

- t - time.
- x - the horizontal dimension, the depth, that depends on time.
- y - the vertical dimension, the height, that also depends on time.

The *path* of the ball will appear to an observer – from the right angle – as a curve. It is placed in the xy-plane positioned vertically:



First, the *sequences*.

We used these *difference quotients* to find the velocity and then the acceleration from the location:

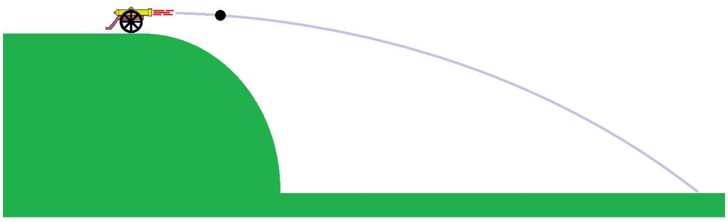
	horizontal	vertical
position	x_n	y_n
velocity	$v_n = \dfrac{x_{n+1} - x_n}{h}$	$u_n = \dfrac{y_{n+1} - y_n}{h}$
acceleration	$a_n = \dfrac{v_{n+1} - v_n}{h}$	$b_n = \dfrac{u_{n+1} - u_n}{h}$

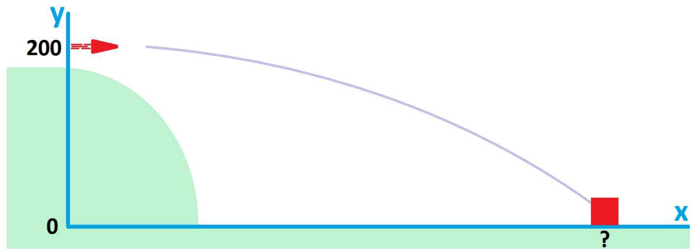
where h is the increment of time.

These formulas can now be solved in order to be able to model the location as a function of time. The result is these recursive formulas for the *Riemann sums*:

	horizontal	vertical
acceleration	a_n	b_n
velocity	$v_{n+1} = v_n + ha_n$	$u_{n+1} = u_n + hb_n$
position	$x_{n+1} = x_n + hv_n$	$y_{n+1} = y_n + hu_n$

Problem: From a 200-feet elevation, a cannon is fired horizontally at 200 feet per second. How far will the cannonball go?





The physics is as follows:

- Horizontal: There is no force, hence $a_n = 0$ for all n.
- Vertical: The force is constant and $b_n = -g$ for all n. Here g is the gravitational constant:

$$g = 32 \text{ ft/sec}^2 .$$

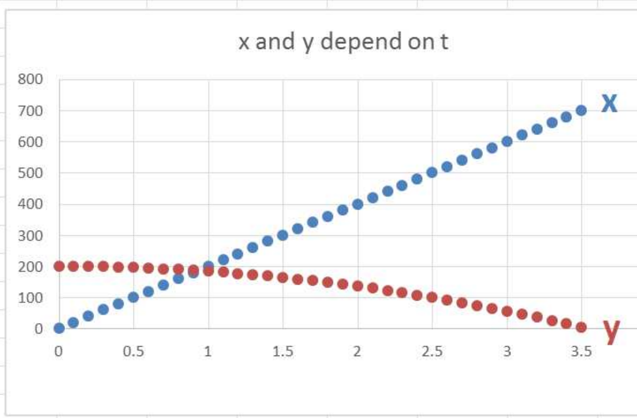
Next, we acquire the initial conditions:

- The initial location: $x_0 = 0$ and $y_0 = 200$.
- The initial velocity: $v_0 = 200$ and $u_0 = 0$.

Example 4.1.1: how far

To find when and where the ball hits the ground, we scroll down to find the row with y closest to 0:

time	depth x	height y
0	0	200
0.1	20	199.84
0.2	40	199.36
0.3	60	198.56
0.4	80	197.44
0.5	100	196
0.6	120	194.24
...	...	...
3.2	640	36.16
3.3	660	25.76
3.4	680	15.04
3.5	700	4
3.6	720	-7.36



It happens sometime between $t = 3.5$ and $t = 3.6$ seconds, say $t_1 = 3.55$ seconds. Second, the values of x during this time period are between $x = 700$ and $x = 720$ feet, say, $x_1 = 710$ feet. We also plot the graphs of x and y as functions of t on the right.

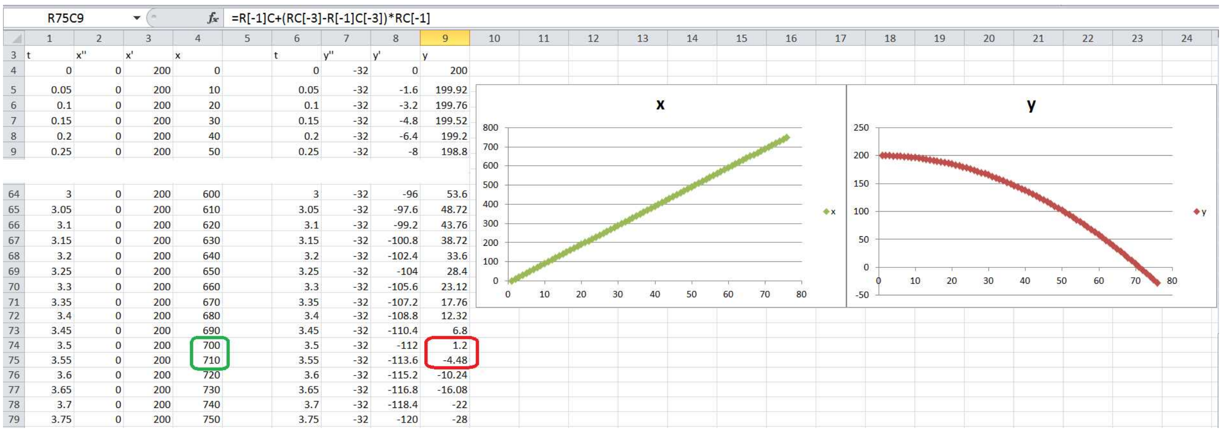
The spreadsheet is constructed for x and y separately, as follows. The time is in the first column progressing from 0 every 0.05. The second derivative is in the next, 0 and -32, respectively. In the next column, the initial velocity is entered in the top cell, 200 and 0 respectively. Below, the velocity is computed as a Riemann sum function of the previous column, with the same formula:

```
=R[-1]C+(RC[-2]-R[-1]C[-2])*R[-1]C[-1]
```

In the next column, the initial location is entered in the top cell, 0 and 200 respectively. Below, the location is computed as a Riemann sum function of the previous column, with the same formula:

```
=R[-1]C+(RC[-3]-R[-1]C[-3])*RC[-1]
```

The results are shown below:

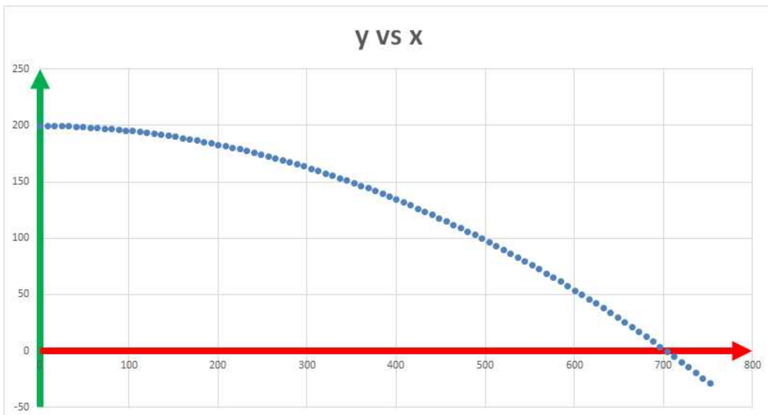


R75C9 =R[-1]C+(RC[-3]-R[-1]C[-3])*RC[-1]

	1	2	3	4	5	6	7	8	9
3	t	x''	x'	x		t	y''	y'	y
4	0	0	200	0		0	-32	0	200
5	0.05	0	200	10		0.05	-32	-1.6	199.92
6	0.1	0	200	20		0.1	-32	-3.2	199.76
7	0.15	0	200	30		0.15	-32	-4.8	199.52
8	0.2	0	200	40		0.2	-32	-6.4	199.2
9	0.25	0	200	50		0.25	-32	-8	198.8
64	3	0	200	600		3	-32	-96	53.6
65	3.05	0	200	610		3.05	-32	-97.6	48.72
66	3.1	0	200	620		3.1	-32	-99.2	43.76
67	3.15	0	200	630		3.15	-32	-100.8	38.72
68	3.2	0	200	640		3.2	-32	-102.4	33.6
69	3.25	0	200	650		3.25	-32	-104	28.4
70	3.3	0	200	660		3.3	-32	-105.6	23.12
71	3.35	0	200	670		3.35	-32	-107.2	17.76
72	3.4	0	200	680		3.4	-32	-108.8	12.32
73	3.45	0	200	690		3.45	-32	-110.4	6.8
74	3.5	0	200	700		3.5	-32	-112	1.2
75	3.55	0	200	710		3.55	-32	-113.6	-4.48
76	3.6	0	200	720		3.6	-32	-115.2	-10.24
77	3.65	0	200	730		3.65	-32	-116.8	-16.08
78	3.7	0	200	740		3.7	-32	-118.4	-22
79	3.75	0	200	750		3.75	-32	-120	-28

To find the solution to the problem from this data, we find the interval during which the cannonball hit the ground, i.e., $y=0$. We go down the y column until we find the value closest to 0; it is $y=1.2$. We then find the corresponding value of x; it is $x=700$.

Plotting x against y produces the path of the cannonball:



Exercise 4.1.2

Under the same conditions, solve numerically the problem of hitting a target 500 feet away.

We start with the *continuous case* now:

- horizontal: $x''=0$
- vertical: $y''=-g$

We start at the same place as above:

$$\begin{aligned} x'' &= 0, & x'(0) &= 200, & x(0) &= 0 \\ y'' &= -g, & y'(0) &= 0, & y(0) &= 200 \end{aligned}$$

Since the velocity is an antiderivative of the acceleration, we integrate these. Then for horizontal, we have:

$$x' = \int 0\, dt = C_x\,,$$

where C_x is any constant. Next, for the vertical,

$$y' = \int -g\, dt = -gt + C_y\,,$$

where C_y is any constant.

Since the location is an antiderivative of the velocity, we integrate these. Then for horizontal, we have:

$$x = \int x'\, dt = \int C_x\, dt = C_x t + K_x\,,$$

where K_x is any constant. Next, for the vertical,

$$y = \int y'\,dx = \int (-gt + C_y)\,dt = -\tfrac{1}{2}gt^2 + C_y t + K_y\,,$$

where K_y is any constant.

Thus, the *general solution* of this system of differential equations is:

$$\begin{array}{lllllll} x & = & & & C_x t & + & K_x\,, \\ y & = & -\frac{1}{2}gt^2 & + & C_y t & + & K_y\,. \end{array}$$

Any possible dynamics is found by specifying the values of the four constants:

$$C_x,\ C_y,\ ,K_x,\ K_y\,.$$

The physics of the situation allows us to assign meanings to these four constants. First,

$$\begin{array}{lllllll} x' & = & & & C_x & \Longrightarrow & x'(0) = C_x\,, \\ y' & = & -gt & + & C_y & \Longrightarrow & y'(0) = C_y\,. \end{array}$$

Therefore,

- C_x is the (constant) horizontal component of velocity;
- C_y is the initial vertical component of velocity.

Next,

$$\begin{array}{lllllllll} x & = & & & C_x t & + & K_x & \Longrightarrow & x(0) = K_x\,, \\ y & = & -\frac{1}{2}gt^2 & + & C_y t & + & K_y & \Longrightarrow & y(0) = K_y\,. \end{array}$$

Therefore,

- K_x is the initial horizontal location (depth);
- K_y is the initial vertical location (height).

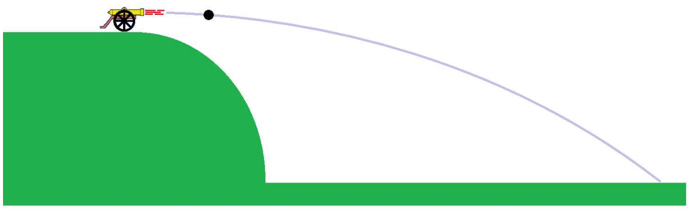
Thus, we have:

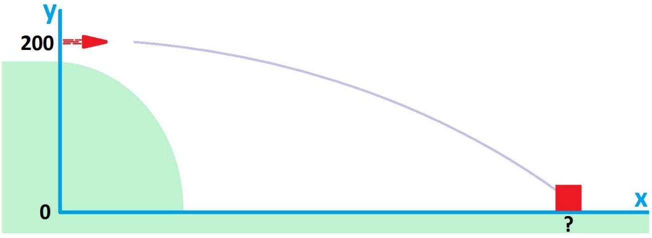
$$\begin{array}{lllllll} \text{Depth} & = & \text{initial depth} & + & \text{initial horizontal velocity} & \cdot\ \text{time} & \\ \text{Height} & = & \text{initial height} & + & \text{initial vertical velocity} & \cdot\ \text{time} & -\frac{1}{2}g\cdot\ \text{time}^{\ 2} \end{array}$$

We used these two equations to solve a variety of problems about motion.

Example 4.1.3: how far

From a 200-feet elevation, a cannon is fired horizontally at 200 feet per second. How far will the cannonball go?





The initial conditions:

- the initial location: 0 and 200.
- the initial velocity: 200 and 0.

Then our equations become:

$$\begin{aligned} x &= \qquad 200t\,, \\ y &= 200 \qquad\quad -16t^2. \end{aligned}$$

Previously we solved the problem *algebraically* as follows. The height at the end of the flight is $y_1 = 0$, so to find the time, we set $y = 200 - 16t^2 = 0$ and solve for t:

$$t_1 = \sqrt{\frac{200}{16}} \approx 3.54\,.$$

We substitute this value of t into x to find the corresponding depth:

$$x_1 = 200t_1 = 200\frac{5\sqrt{2}}{2} \approx 707\,.$$

What about the *velocity* as a function of time? We have:

$$\begin{aligned} \frac{dx}{dt} &= v_x\,, \\ \frac{dy}{dt} &= v_y \quad -gt\,. \end{aligned}$$

Adding these two equations to the former two allows us to solve more profound problems.

Example 4.1.4: impact

In the setting of the last example, how *hard* does the ball hit the ground?

First, we examine the spreadsheet. Instead of the formulas, we compute the average velocities (i.e., the difference quotients) to approximate the velocities. The formula for x' is:

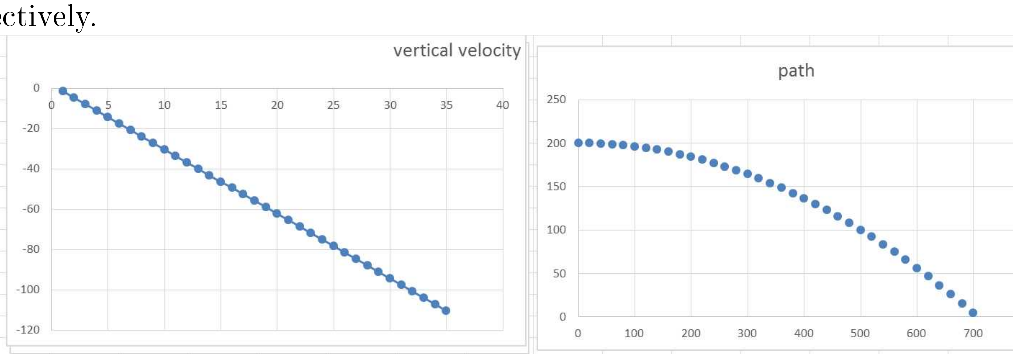
```
=(RC[-2]-R[-1]C[-2])/(RC[-3]-R[-1]C[-3])
```

and the formula for y' is:

```
=(RC[-2]-R[-1]C[-2])/(RC[-4]-R[-1]C[-4])
```

The denominators refer to the column that contains the time, and the numerator refers to the columns that contain x and y respectively.

time	depth x	height y	x'	y'
0	0	200		
0.1	20	199.84	200	-1.6
0.2	40	199.36	200	-4.8
0.3	60	198.56	200	-8
0.4	80	197.44	200	-11.2
0.5	100	196	200	-14.4
....				
3.5	700	4	200	-110.4
3.6	720	-7.36	200	-113.6



Looking at the same row as before, we see that the vertical velocity at the moment of impact is between -110.4 and -113.6 feet per second.

Now, the algebra. The formulas for the velocities take this form:

$$\begin{aligned} \frac{dx}{dt} &= 200\,, \\ \frac{dy}{dt} &= \qquad -32t\,. \end{aligned}$$

Let's find the velocity at the time of contact. We substitute the time we've found,

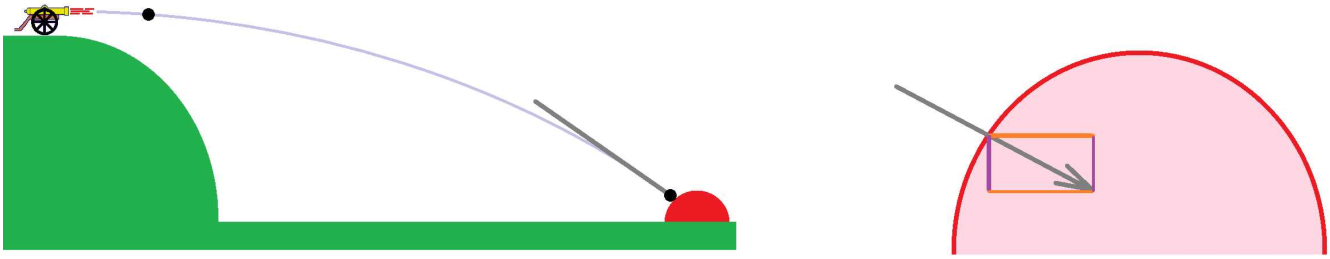
$$t_1 = \frac{5\sqrt{2}}{2}\,,$$

into the formulas for velocity:

$$\begin{aligned}\frac{dx}{dt}\Big|_{t=t_1} &= 200,\\ \frac{dy}{dt}\Big|_{t=t_1} &= \quad -32t_1 = -32\frac{5\sqrt{2}}{2} \approx -112\,.\end{aligned}$$

The answer matches our estimate.

But which one of the two numbers represent how *fast* the ball hits the ground? It is the latter if the ball hits the (horizontal) surface, and it is the former if this is a wall. Then, the general answer should be a combination of the two. This is how they should be combined via the *Pythagorean Theorem*:



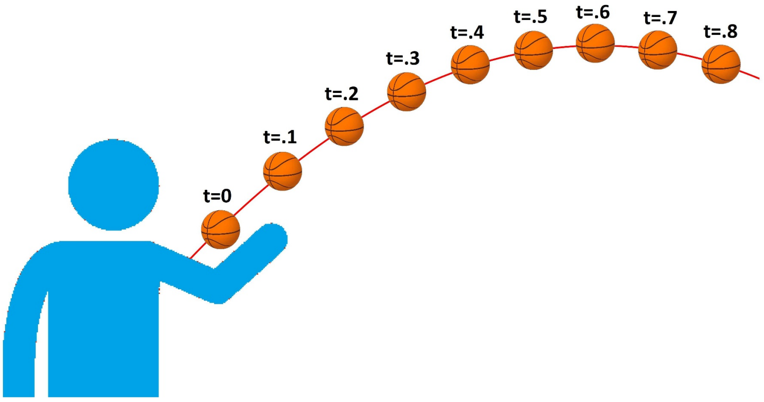
Then, the impact is determined by this number:

$$\sqrt{200^2 + (-112)^2} \approx 229\,.$$

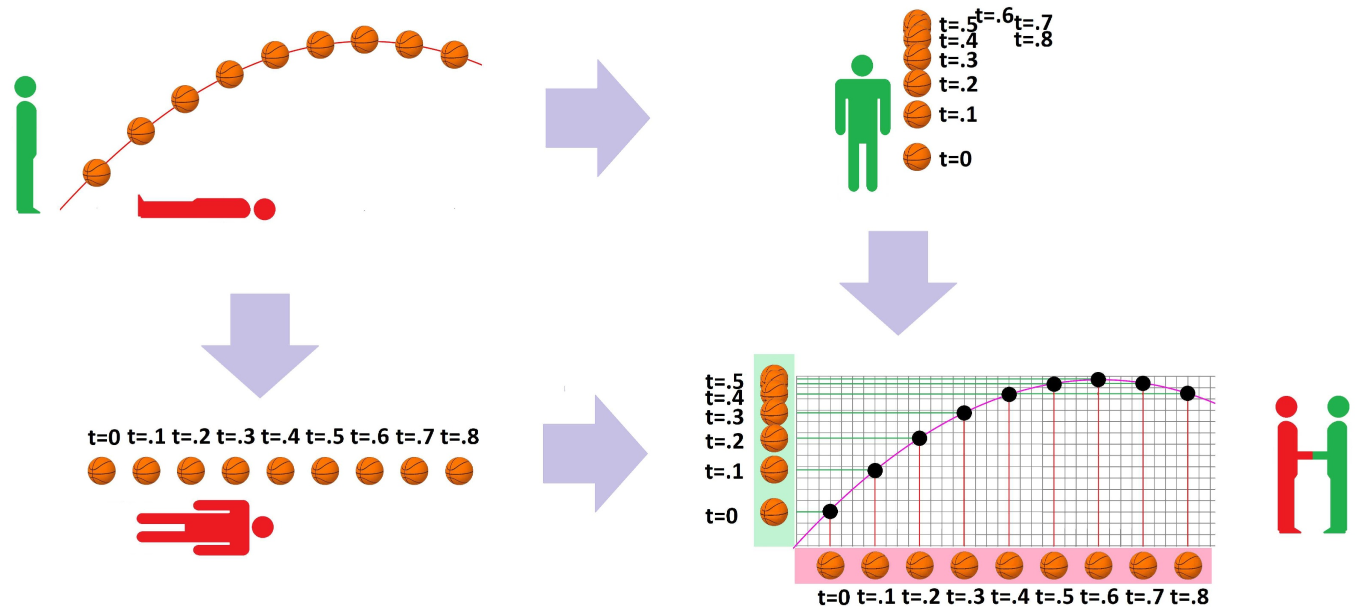
4.2. Parametric curves on the plane

Example 4.2.1: ball

Imagine a person observing the flight of a thrown ball as it passes by:



Is there another way to capture this flight? Imagine there are two more observers:
- The first one (red) is on the ground under the path of the ball and can only see the forward progress of the ball.
- The second one (green) is behind the throw and can see only the rise and fall of the ball.



If the two make records of where the ball was at what time, they can use the *time stamps* to match the two coordinates and then plot this point on the xy-plane. These points will form the ball's trajectory, what the first observer saw. It is called a *parametric curve*.

Curves aren't represented as graphs of functions. In fact, y doesn't depend on x anymore, but they are *related* to each other. The link is established by means of another variable, t. So, we have two functions that have nothing to do with each other except the inputs can be matched.

Definition 4.2.2: parametric curve

A *parametric curve* on the plane is a combination of two functions of the same variable:

$$\boxed{\begin{cases} x = f(t) \\ y = g(t) \end{cases}}$$

We can also use the Cartesian system of points to represent this curve:

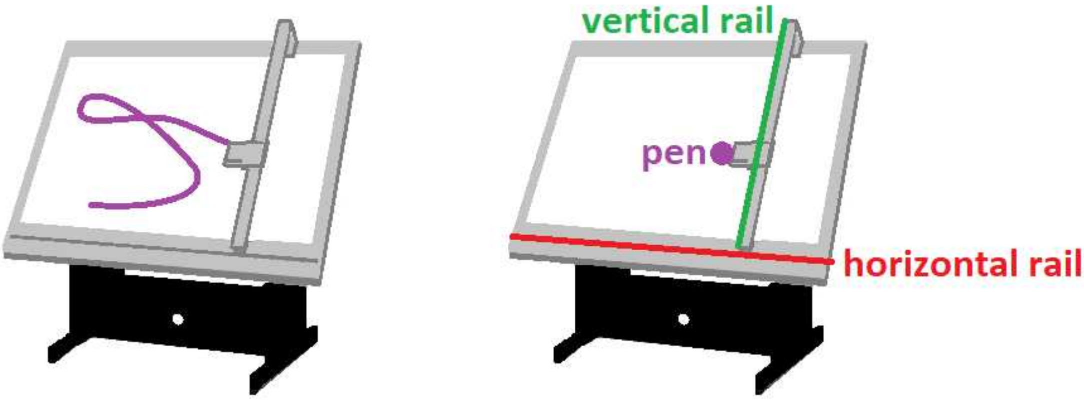
$$\boxed{(x, y) = \big(f(t), g(t)\big)}$$

Exercise 4.2.3

Explain how a parametric curve is a relation.

Example 4.2.4: plotter

A curve may be plotted on a piece of paper by hand or by a computer by the following method. A pen is attached to a runner on a vertical bar, while that bar slides along a horizontal rail at the bottom edge of the paper:



The computer commands the next location of both as follows. At each moment of time t, we have:

1. The horizontal location of the vertical bar (and the pen) is given by $x = f(t)$.
2. The vertical location of the pen is given by $y = g(t)$.

Warning!

This view of parametric curves is most useful within the framework of multidimensional spaces and vectors.

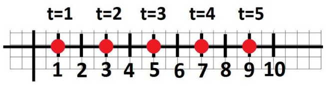
Example 4.2.5: straight lines

Let's examine motion along a straight line on the xy-plane.

First we go along the x-axis. The motion is represented by a familiar linear function of time:

$$x = 2t + 1\,.$$

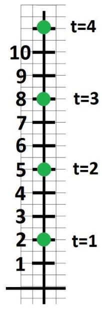
We are moving 2 feet per second to the right starting at $x = 1$. These are a few of the locations:



Second we go along the y-axis. The motion is represented by another linear function of time:

$$y = 3t + 2\,.$$

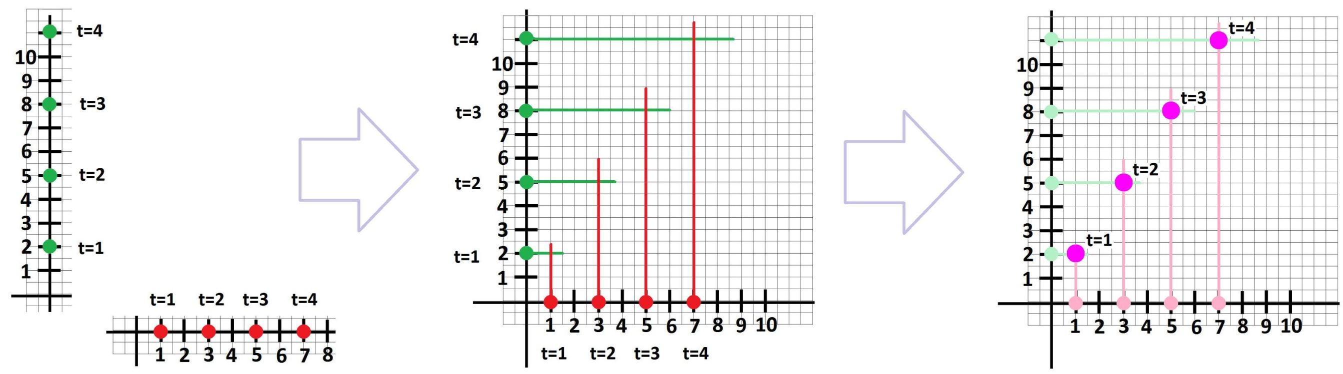
We are moving 3 feet per second up starting at $y = 2$. These are a few of the locations:



Now, what if these two are just two different views of the same motion from two different observers? Then we have:

$$\begin{cases} x = 2t + 1\,, \\ y = 3t + 2\,. \end{cases}$$

These are a few of the locations:



Exercise 4.2.6

Explain why these points lie on a straight line. Hint: triangles.

Of course, the motion metaphor – x and y are coordinates in the space – will be superseded. In contrast to this approach, we look at the two quantities and two functions that might have *nothing* to do with each other (except for t, of course).

Example 4.2.7: commodities trader

Suppose a commodities trader follows the market. What he sees is the following:

- t - time
- x - the price of wheat (say, in dollars per bushel)
- y - the price of sugar (say, in dollars per ton)

We simply have two functions and we – initially – look at them separately.

First, let's imagine that the price of *wheat* is decreasing:

$$x \searrow$$

The data comes to the observer in a pure, numerical form. To simulate this situation and to make this specific, one can choose a formula, for example:

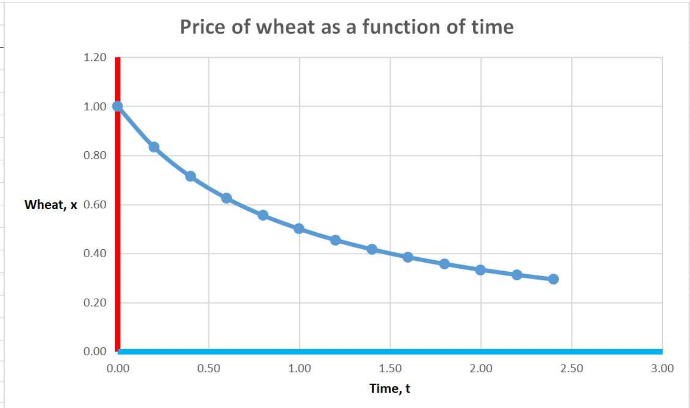
$$x = f(t) = \frac{1}{t+1}\,.$$

To show some actual data, we evaluate x for several values of t:

t	x
0	1.00
1	0.50
2	0.33

With more points acquired in a spreadsheet, we can plot the graph on the tx-plane:

Input, time t	Output, price of wheat x
0.00	1.00
0.20	0.83
0.40	0.71
0.60	0.63
0.80	0.56
1.00	0.50
1.20	0.45
1.40	0.42
1.60	0.38
1.80	0.36
2.00	0.33
2.20	0.31
2.40	0.29



At this point, we could, if needed, apply the available apparatus to study the symmetries, the monotonicity, the extreme points, etc. of this function.

Second, suppose that the price of *sugar* is increasing and then decreasing:

$$y \nearrow \searrow$$

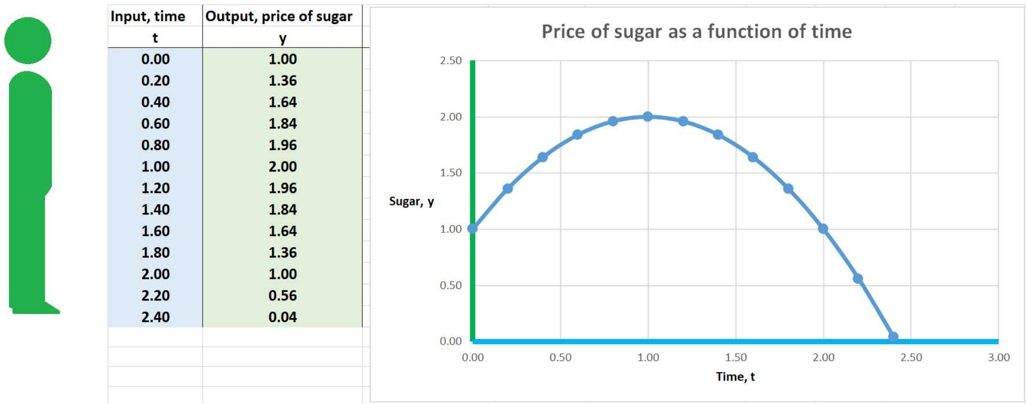
To make this specific, we can choose an upside-down parabola:

$$y = g(t) = -(t-1)^2 + 2\,.$$

We then again evaluate y for several values of t:

t	y
0	1.00
1	2.00
2	1.00

With more points acquired in a spreadsheet, we plot the graph on the ty-plane:



Input, time t	Output, price of sugar y
0.00	1.00
0.20	1.36
0.40	1.64
0.60	1.84
0.80	1.96
1.00	2.00
1.20	1.96
1.40	1.84
1.60	1.64
1.80	1.36
2.00	1.00
2.20	0.56
2.40	0.04

What if the trader is interested in finding hidden relations between these *two commodities.* Let's combine the data first:

t	x
0	1.00
1	0.50
2	0.33

and

t	y
0	1.00
1	2.00
2	1.00

$\longrightarrow$

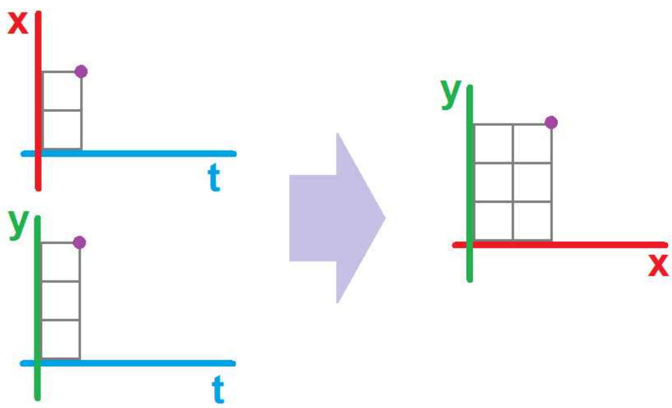
t	x	y
0	1.00	1.00
1	0.50	2.00
2	0.33	1.00

Since the input t is the same, we give it a single column. There seems to be two outputs. A better idea is to see *pairs* (x, y):

t	$($ x , y $)$
0	(1.00 , 1.00)
1	(0.50 , 2.00)
2	(0.33 , 1.00)

A value of x is paired up with a value of y when they appear along the same t in both plots.

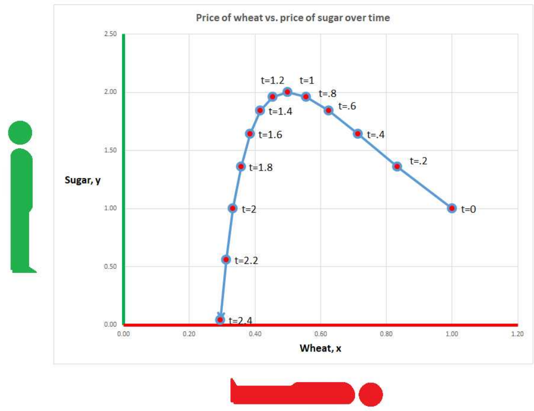
How do we combine the two *plots*? As the two plots are made of (initially) disconnected points – (t, x) and (t, y) – so is the new plot. This is what happens to each pair:



There is no t! As the independent variable is the same for both functions, only the dependent variables appear. Instead of plotting all points (t, x, y), which belong to the 3-dimensional space, we just plot (x, y) on the xy-plane – for each t. It's a "scatter plot" connected to make a curve:

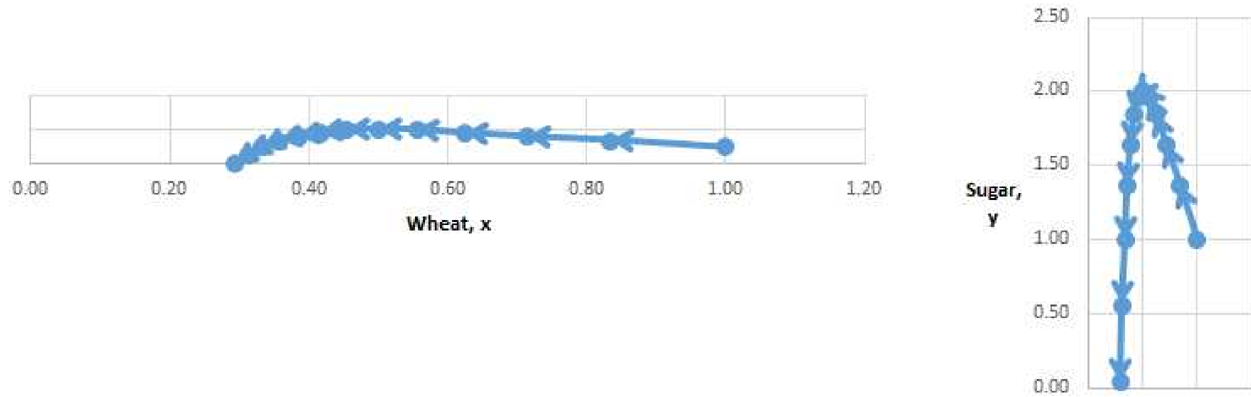
Input, time	Output, price of wheat	Input, time	Output, price of sugar
t	x	t	y
0.00	1.00	0.00	1.00
0.20	0.83	0.20	1.36
0.40	0.71	0.40	1.64
0.60	0.63	0.60	1.84
0.80	0.56	0.80	1.96
1.00	0.50	1.00	2.00
1.20	0.45	1.20	1.96
1.40	0.42	1.40	1.84
1.60	0.38	1.60	1.64
1.80	0.36	1.80	1.36
2.00	0.33	2.00	1.00
2.20	0.31	2.20	0.56
2.40	0.29	2.40	0.04

The direction matters! Since t is missing, we have to make sure we know in which direction we are moving and indicate that with an arrow. Ideally, we also label the points in order to indicate not only "where" but also "when":



Thus, this is motion, just as before, but through what space? An abstract *space of prices* that we've made up. The space is comprised of all possible combinations of prices, i.e., a point (x, y) stands for a combination of two prices: x for wheat and y for sugar.

How much information about the dynamics of the two prices contained in the original functions can we recover from the new graph? A lot. We can shrink the graph vertically to de-emphasize the change of y and to reveal the *qualitative* behavior of x, and vice versa:



We see the decrease of x and then the increase followed by the decrease of y. In addition, the density of the points indicates the speed of the motion.

Example 4.2.8: abstract

We can do this in a fully abstract setting. When two functions, f, g, are represented by their respective lists of values (instead of formulas), they are easily combined into a parametric curve, F. We just need to eliminate the repeated column of inputs. Suppose we need to combine these two functions:

$$\begin{array}{c|c} t & x=f(t) \\ \hline 0 & 1 \\ 1 & 2 \\ 2 & 3 \\ 3 & 0 \\ 4 & 1 \end{array} \quad \& \quad \begin{array}{c|c} t & y=g(t) \\ \hline 0 & 5 \\ 1 & -1 \\ 2 & 2 \\ 3 & 3 \\ 4 & 0 \end{array} \quad = \quad ?$$

We repeat the inputs column – only once – and then repeat the outputs of either function. First row:

$$f: 0 \mapsto 1 \quad \& \quad g: 0 \mapsto 5 \quad \Longrightarrow \quad F: 0 \mapsto (0, 5)$$

Second row:

$$f: 1 \mapsto 2 \quad \& \quad g: 1 \mapsto -1 \quad \Longrightarrow \quad F: 1 \mapsto (2, -1)$$

And so on. This is the whole solution:

$$\begin{array}{c|c} t & x=f(t) \\ \hline 0 & 1 \\ 1 & 2 \\ 2 & 3 \\ 3 & 0 \\ 4 & 1 \end{array} \quad \text{and} \quad \begin{array}{c|c} t & y=g(t) \\ \hline 0 & 5 \\ 1 & -1 \\ 2 & 2 \\ 3 & 3 \\ 4 & 0 \end{array} \quad \longrightarrow \quad \begin{array}{c|lcr} t & P= \ (f(t) & , & g(t)) \\ \hline 0 & (1 & , & 5) \\ 1 & (2 & , & -1) \\ 2 & (3 & , & 2) \\ 3 & (0 & , & 3) \\ 4 & (1 & , & 0) \end{array}$$

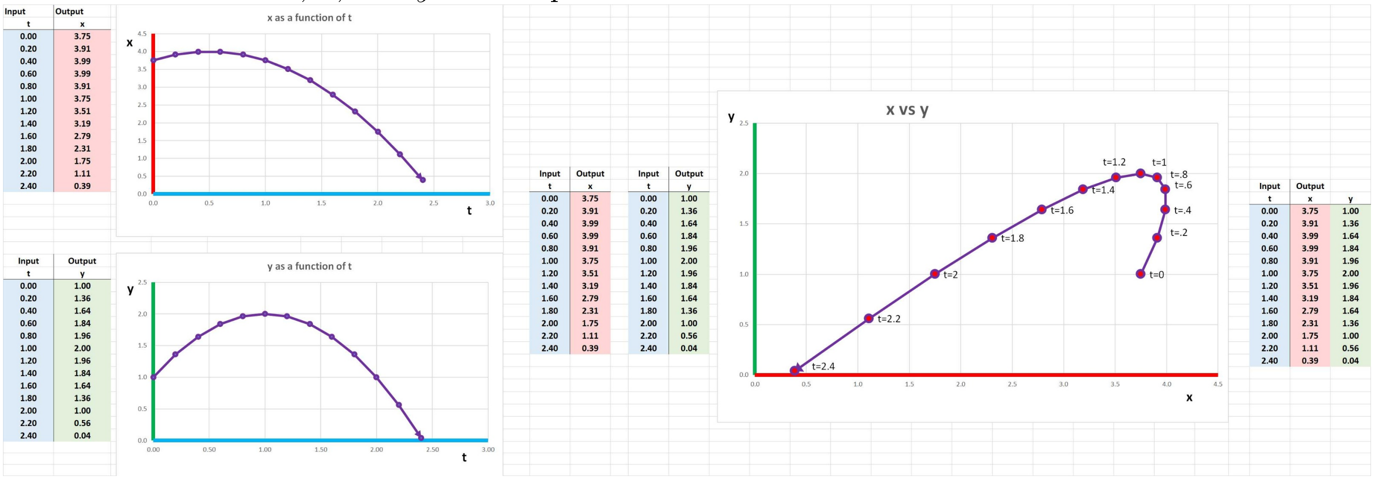
As you can see, there are no algebraic operations carried out and there is no new data, just the old data arranged in a new way. However, it is becoming clear that the list is also a function of some kind.

Warning!

The end result isn't the graph of any function.

Example 4.2.9: spreadsheet

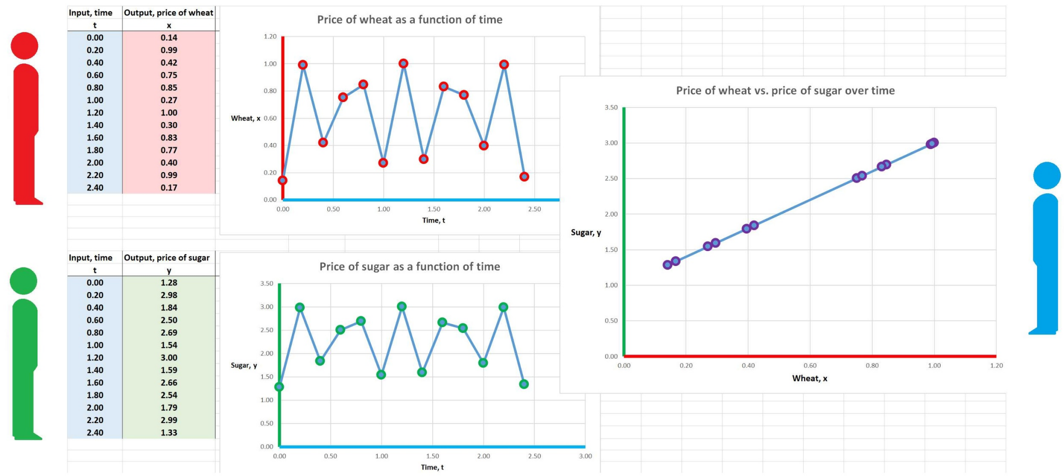
This is a summary of how the parametric curve is formed from two functions provided in a spreadsheet. The three columns – t, x, and y – are copied and then the last two are used to create a chart:



This chart is the *path* – not the graph – of the parametric curve. Note also that the curve isn't the graph of any function of one variable as the *Vertical Line Test* is violated.

Example 4.2.10: pattern

Plotting a parametric curve may reveal a relation between two quantities:



Parametric curves are functions!

This idea comes with certain obligations (Volume 1). First, we have to *name* it, say F. Second, as we combine the two functions, we use the following notation for this operation:

parametric curve

$$F = (f, g) : \begin{cases} x = f(t) \\ y = g(t) \end{cases}$$

Next, what is the *independent variable*? It is t. After all, this is the input of both of the functions involved. What is the *dependent variable*? It is the "combination" of the outputs of the two functions, i.e., x and y. We know how to combine these; we form a pair, $P = (x, y)$. This P is a point on the xy-plane!

To summarize, we do what we have done many times before (addition, multiplication, etc.) – we create a new function from two old functions. We represent a function f diagrammatically as a *black box* that

processes the input and produces the output:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output} \\ t & \to & \boxed{\quad f \quad} & \to & x \end{array}$$

Now, what if we have another function g:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output} \\ t & \to & \boxed{\quad g \quad} & \to & y \end{array}$$

How do we represent $F = (f, g)$? To represent it as a single function, we need to "wire" their diagrams together side by side:

$$\begin{array}{ccccc} t & \to & \boxed{\quad f \quad} & \to & x \\ || & & & & \updownarrow \\ t & \to & \boxed{\quad g \quad} & \to & y \end{array}$$

It is possible because the input of f is the same as the input of g. For the outputs, we can combine them even when they are of different nature. Then we have a diagram of a new function:

$$(f,g):\ t\ \to\ \boxed{\begin{array}{ccccccccc} & \nearrow & t & \to & \boxed{\quad f \quad} & \to & x & \searrow & \\ t & & & & & & & & (x,y) \\ & \searrow & t & \to & \boxed{\quad g \quad} & \to & y & \nearrow & \end{array}}\ \to\ P$$

We see how the input variable t is copied into the two functions, processed by them *in parallel*, and finally the two outputs are combined together to produce a single output. The result can be seen again as a black box:

$$t\ \to\ \boxed{\quad F \quad}\ \to\ P$$

The difference from all the functions we have seen so far is the nature of the output.

What about the *image* (the range of values) of $F = (f, g)$? It is supposed to be a recording of all possible outputs of F. The terminology used is often different though.

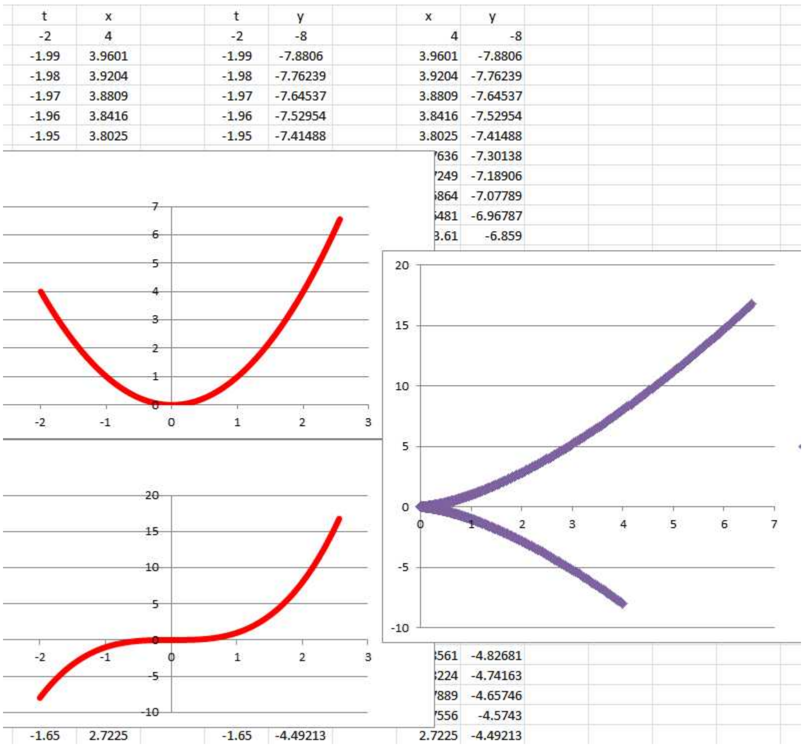
Definition 4.2.11: path of parametric curve

The *path* of a parametric curve $x = f(t),\ y = g(t)$ is the set of all such points $P = (f(t), g(t))$ on the xy-plane.

The path is typically a curve. We plot several of them below.

Example 4.2.12: path

In general, the two processes, $x = x(t)$ and $y = y(t)$, are independent. When we combine them to see the path of the object by plotting (x, y) for each t, the result may be unexpected:



What about the *graph* of $F=(f,g)$? As we know from Chapter 1, the graph of a function is supposed to be a recording of all possible combinations of inputs and outputs of F. What if the outputs are 2-dimensional?

Definition 4.2.13: graph of parametric curve

The *graph* of a parametric curve $x=f(t),\ y=g(t)$ is the set of all points of the form:

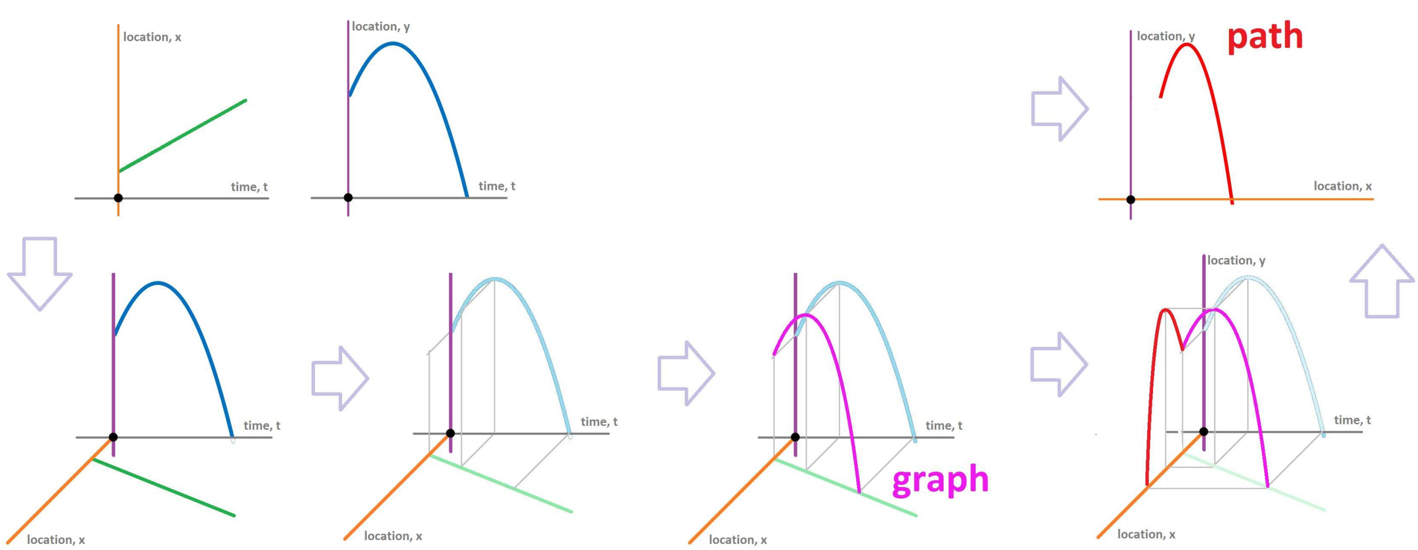
$$(t,x,y)=(t,f(t),g(t))$$

in the txy-space.

The graph is built from these two:

- the graph of $x=f(t)$ on the tx-plane (the floor), and
- the graph of $y=g(t)$ on the ty-plane (the wall facing us).

It is a curve in space, akin to a piece of wire:



Then the shadow of this wire on the floor is the graph $x=f(t)$ (light from above). If the light is behind us, the shadow on the wall in front is the graph $y=g(t)$. In addition, pointing a flashlight from right to left

will produce the path of the parametric curve on the xy-plane.

This is the summary of the terminology:

types of functions:	general functions	numerical functions	parametric curves	motion
the set of all outputs:	image	range	path	trajectory

4.3. Functions of two variables

Any formula with two independent variables and one dependent variable can be studied in this manner:

$$a = wd \text{ or } z = x + y\,.$$

Such an expression is called a *function of two variables.* The notation is as follows:

$$g(w, d) = wd \text{ or } f(x, y) = x + y\,.$$

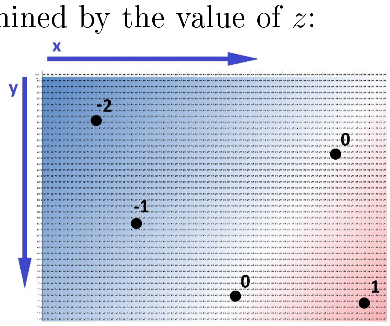
Example 4.3.1: function of two variables

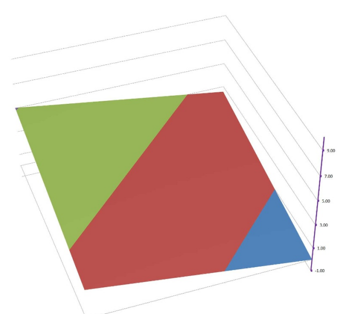
Let

$$f(x, y) = x + y\,.$$

We illustrate this new function below. First, by changing – independently — the two variables, we create a table of numbers (left). We can, furthermore, color this array of cells (middle) so that the color of the (x, y)-cell is determined by the value of z:

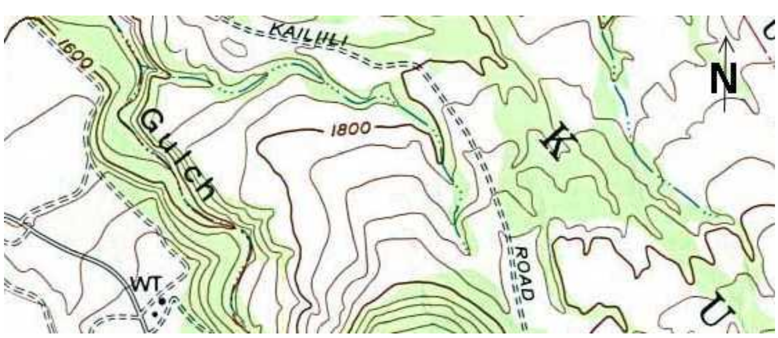
x\y	0	10	20	30	40	50	60
0	0	10	20	30	40	50	60
10	10	20	30	40	50	60	70
20	20	30	40	50	60	70	80
30	30	40	50	60	70	80	90
40	40	50	60	70	80	90	100
50	50	60	70	80	90	100	110
60	60	70	80	90	100	110	120
70	70	80	90	100	110	120	130
80	80	90	100	110	120	130	140





The value of z can also be visualized as the elevation of a point at that location (right).

So, the main metaphor for a function of two variables will be *terrain*:



Each line indicates a constant elevation.

Example 4.3.2: distance

The distance formula for the Cartesian plane creates a function of two variables. This is the distance from a point (x, y) to the origin:

$$z = \sqrt{x^2 + y^2}\,.$$

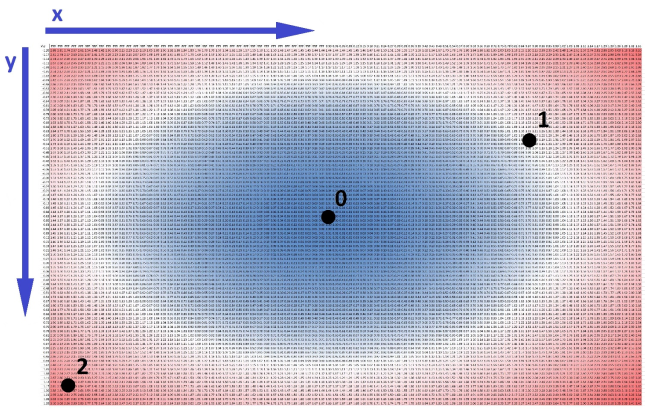
Slightly simpler is the square of the distance from a point (x, y) to the origin:

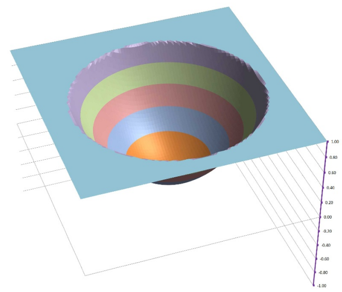
$$z = x^2 + y^2 .$$

We create a table of the values of the expression on the left in a spreadsheet with the formula:

```
=RC1^2+R1C^2
```

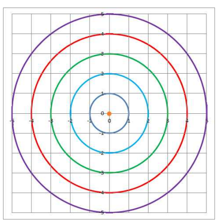
We then color the cells:





The negative values of z are in blue and the positive are in red. The circular pattern is clear.

The pattern seems to be made from concentric circles with the radii that vary with z:



For each z, we have a *relation* between x and y.

We also represent a function p diagrammatically as a *black box* that processes the inputs and produces the output:

$$\begin{array}{ccccc} \text{inputs} & & \text{function} & & \text{output} \\ x & & & & \\ & \searrow & & & \\ & & \boxed{\ p\ } & \to & z \\ & \nearrow & & & \\ y & & & & \end{array}$$

Instead, we would like to see a single input variable, (x, y), decomposed into two, x and y, to be processed by the function *at the same time*:

$$(x, y) \ \to \ \boxed{\ p\ } \ \to \ z$$

The difference from all the functions we have seen so far is the nature of the input.

So, even though we speak of *two* variables, the idea of function remains the same:

- ▶ There is a set (domain) and another (codomain) and the function assigns to each element of the former an element of the latter.

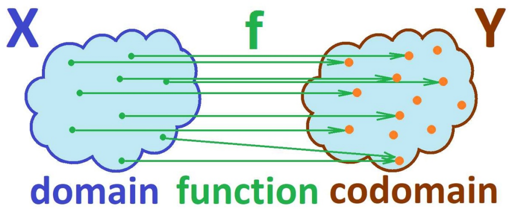
The idea is reflected in the notation we use:

$$F : X \to Z$$

or

$$X \xrightarrow{\ F\ } Z$$

A common way to visualize the concept of function – especially when the sets cannot be represented by mere lists – is to draw shapeless blobs connected by arrows:



In contrast to numerical function, however, *the domain is a subset of the* (x, y)*-plane.*

For example, we have for $f(x, y) = x + y$:

$$(0,0) \to 0,\ (0,1) \to 1,\ (1,0) \to 1,\ (1,1) \to 2,\ (1,2) \to 3,\ (2,1) \to 3,\ ...$$

Each arrow clearly identifies the *input* (an element of X) of this procedure by its beginning and the *output* (an element of Z) by its end.

This is the notation for the output of a function F when the input is x:

Input and output of function

$$F(x, y) = z$$

or

$$F : (x, y) \ \to \ z$$

It reads: "F of (x, y) is z".

We still have:

$$F(\text{ input }) = \text{ output}$$

and

$$F : \text{ input } \ \to \ \text{ output }.$$

Functions are *explicit relations.* There are *three* variables related to each other, but this relation is unequal: The two input variables come first and, therefore, the output is *dependent* on the input. That is why we say that the inputs are the *independent variables* while the output is the *dependent variable.*

Example 4.3.3: flowcharts represent functions

For example, for a given input (x, y), we first split it: x and y are the two *numerical* inputs. Then we do the following consecutively:

- add x and y,
- multiply by 2, and then
- square.

Such a procedure can be conveniently visualized with a "flowchart":

$$(x, y) \ \to \ \boxed{\ x + y\ } \ \to \ u \ \to \ \boxed{\ u \cdot 2\ } \ \to \ z \ \to \ \boxed{\ z^2\ } \ \to \ v$$

Functions of two variables come from many sources and can be expressed in different forms:

- a list of instructions (an algorithm)
- an algebraic formula
- a list of pairs of inputs and outputs
- a graph

- a transformation

An *algebraic representation* is exemplified by $z = x^2y$. In order to properly introduce this as a function, we give it a name, say f, and write:

$$f(x, y) = x^2y\,.$$

Let's examine this notation:

Function of two variables

	z	$=$	f	$($	$x\ ,\ y$	$)$	$=$	x^2y
	$\uparrow$		$\uparrow$		$\uparrow\ \uparrow$			$\uparrow\uparrow$
name:	dependent variable		function		independent variables			independent variables

Example 4.3.4: plug in values

Insert one input value in all of these boxes and the other in those circles. For example, this function:

$$f\,(x) = \frac{2x^2y - 3y + 7}{y^3 + 2x + 1}\ ,$$

can be understood and evaluated via this diagram:

$$f\,(\square) = \frac{2\square^2 \bigcirc - 3 \bigcirc + 7}{\bigcirc^3 + 2\square + 1}\,.$$

This is how $f(3, 0)$ is evaluated:

$$f\left(\boxed{3}, 0\right) = \frac{2\boxed{3}^2 0 - 3 0 + 7}{0^3 + 2\boxed{3} + 1}\,.$$

In summary,

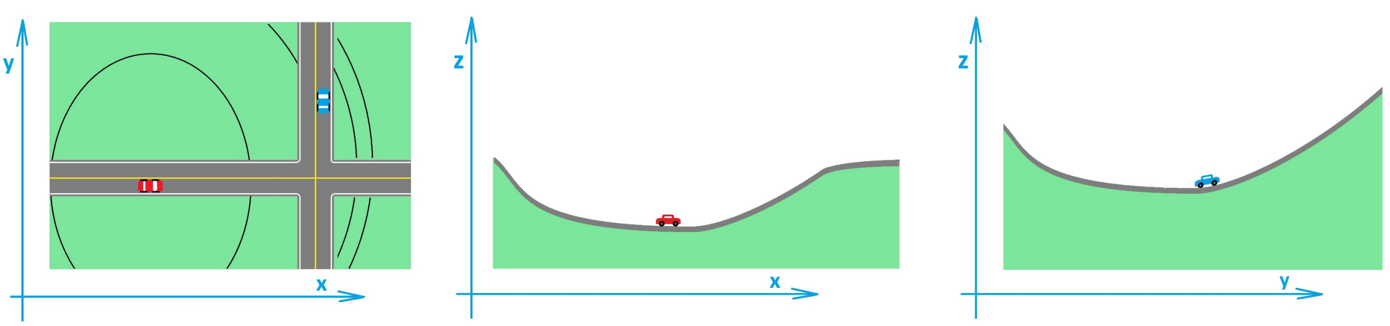
- ▶ "x" and "y" in a formula serve as *placeholders* for: numbers, variables, and whole functions.

How do we study a function of two variables? We use what we know about functions of single variable.

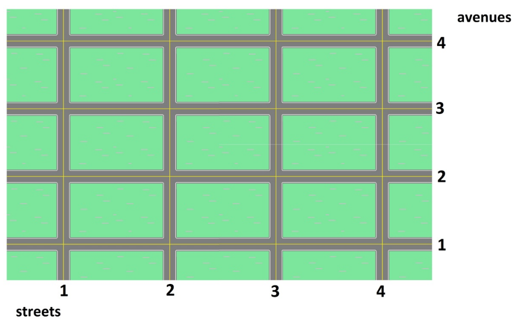
Above we looked at the curves of constant elevation of the surfaces. An alternative idea is a surveying method:

- ▶ In order to study a terrain, we concentrate on the two main directions.

Imagine that we drive south-north and east-west separately, watching how the elevation changes:

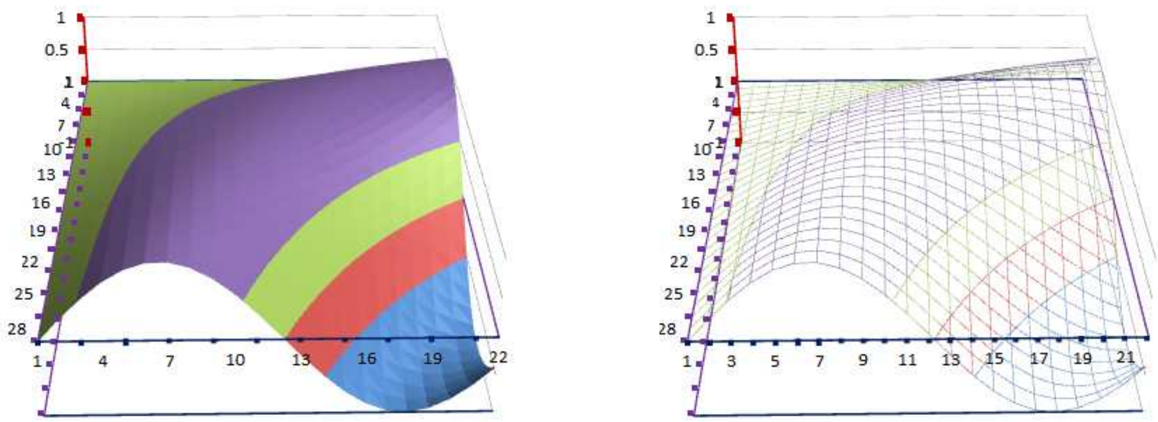


We can even imagine that we drive around a city on a hill and these trips follow the street grid:



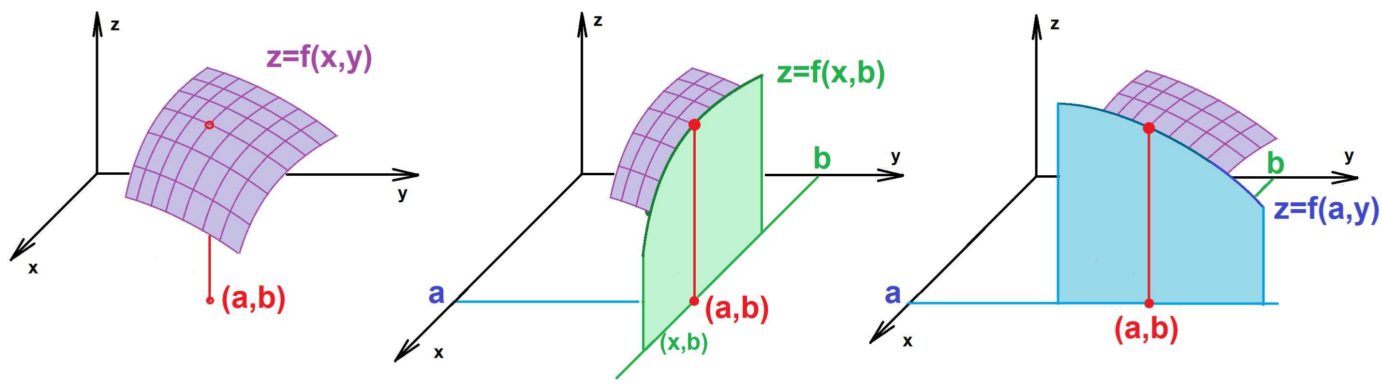
Each of these trips creates a function of single variable, x or y.

To visualize, consider the plot of $F(x,y) = \sin(xy)$ on the left:



We plot the surface as a "wire-frame" on the right. Each wire is a separate trip.

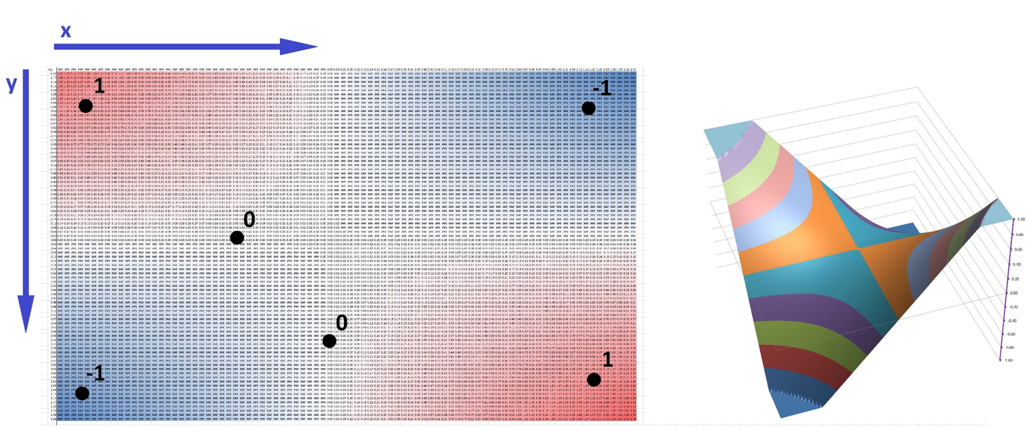
The graphs of these functions are the slices cut by the vertical planes aligned with the axes from the surface that is the graph of F:



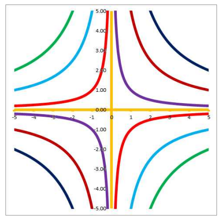
Example 4.3.5: saddle

Let's plot the graph of the function:

$$z = xy\,.$$



This is what the graphs of these relations look like when plotted for various z's; they are curves called *hyperbolas*:

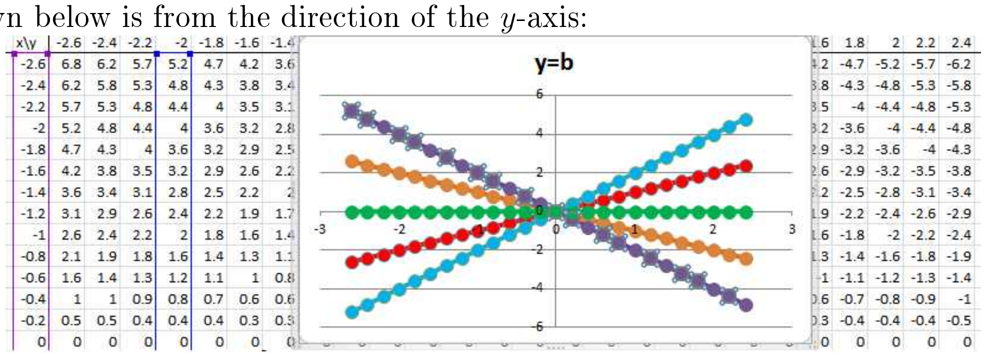


Instead, we fix one independent variable at a time.

We fix y first:

plane	equation	curve
$y = 2$	$z = x \cdot 2$	line with slope 2
$y = 1$	$z = x \cdot 1$	line with slope 1
$y = 0$	$z = x \cdot 0 = 0$	line with slope 0
$y = -1$	$z = x \cdot (-1)$	line with slope 1
$y = -2$	$z = x \cdot (-2)$	line with slope -2

The view shown below is from the direction of the y-axis:

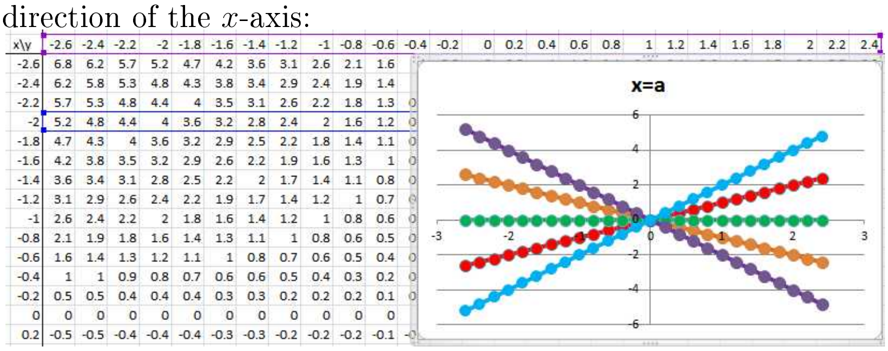


The data for each line comes from the x-column of the spreadsheet and one of the z-columns. These lines give the lines of elevation of this terrain in a particular, say, east-west direction. This is equivalent to cutting the graph by a vertical plane parallel to the xz-plane.

We fix x second:

plane	equation	curve
$x = 2$	$z = 2 \cdot y$	line with slope 2
$x = 1$	$z = 1 \cdot y$	line with slope 1
$x = 0$	$z = 0 \cdot y = 0$	line with slope 0
$x = -1$	$z = (-1) \cdot y$	line with slope 1
$x = -2$	$z = (-2) \cdot y$	line with slope -2

This is equivalent to cutting the graph by a vertical plane parallel to the yz-plane. The view shown below is from the direction of the x-axis:



The data for each line comes from the y-row of the spreadsheet and one of the z-rows. These lines give the lines of elevation of this terrain in a particular, say, north-south direction.

Exercise 4.3.6

Provide a similar analysis for $f(x,y)=3x+2y$.

Example 4.3.7: baker

We will take a look at the example in the last section from a different angle. The time t is not a part of our consideration anymore but we retain the two variables representing the two *commodities*:

- x is the price of wheat.
- y is the price of sugar.

We also add a *product* to the setup:

- z is the price of a loaf of bread.

What is the relation between these three? As those two are the two major ingredients in bread, we will assume that

▶ z depends on x and y.

One can imagine a baker who every morning, upon receiving the updated prices of wheat and sugar, uses a *table* that he made up in advance to decide on the price of his bread for the rest of the day. Let's see how such a table might come about.

What kind of dependencies are these? Increasing prices of the ingredients increases the cost and ultimately the price of the product:

$$x \nearrow \Longrightarrow z \nearrow$$
$$y \nearrow \Longrightarrow z \nearrow$$

At its simplest, such an increase is linear. In addition to some fixed costs:

- Each increase of x leads to a proportional increase of z.
- Each increase of y leads to a proportional increase of z.

Independently! A simple formula that captures this dependence may be this:

$$z=p(x,y)=2x+y+1\,.$$

In order to visualize this function, we compute a few of its values:

- $p(0,0)=1$
- $p(0,1)=2$
- $p(0,2)=3$
- $p(1,0)=3$
- $p(1,1)=4$
- etc.

Even though this is a list, we realize that the input variables don't fit into a list comfortably – they form a *table*!

$$\begin{array}{cccc}(0,0) & (1,0) & (2,0) & ...\\ (0,1) & (1,1) & (2,1) & ...\\ (0,2) & (1,2) & (2,2) & ...\\ ... & ... & ... & ...\end{array}$$

In fact, we can align these pairs with x in each column and y in each row:

$y\backslash x$	0	1	2	...
0	$(0,0)$	$(1,0)$	$(2,0)$	...
1	$(0,1)$	$(1,1)$	$(2,1)$	...
2	$(0,2)$	$(1,2)$	$(2,2)$	...
...	...	...	...	...

Now, the values, $z = p(x, y)$:

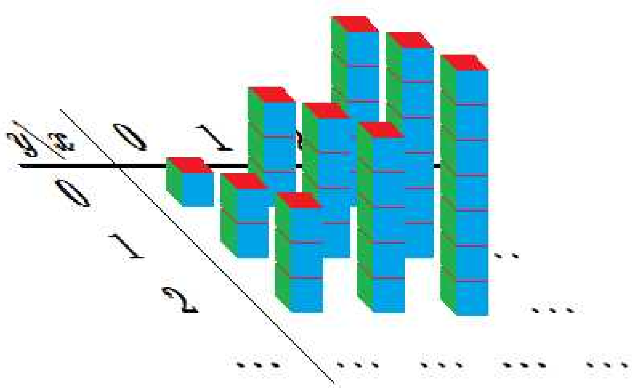
$y\backslash x$	0	1	2	...
0	1	3	5	...
1	2	4	6	...
2	3	5	7	...
...	...	...	...	...

That's what the baker's table might look like.

Let's bring these two together:

$y\backslash x$	0		1		2		...
0	$(0,0)$		$(1,0)$		$(2,0)$		...
		$\searrow$		$\searrow$		$\searrow$	...
		1		3		5	...
1	$(0,1)$		$(1,1)$		$(2,1)$		...
		$\searrow$		$\searrow$		$\searrow$	...
		2		4		6	...
2	$(0,2)$		$(1,2)$		$(2,2)$		...
		$\searrow$		$\searrow$		$\searrow$	...
		3		5		7	...
...	...		...		...		...

In the past, we have visualized numerical functions by putting *bars* on top of the x-axis. Now, we visualize the values by building *columns* with appropriate heights on top of the xy-plane:



Notice that by fixing one of the variables – $x = 0, 1, 2$ or $y = 0, 1, 2$ – we create a function of *one* variable with respect to the other variable. We fix x below and extract the *columns* from the table:

$x = 0$:

y	z
0	1
1	2
2	3

$x = 1$:

y	z
0	3
1	4
2	5

$x = 2$:

y	z
0	5
1	6
2	7

A *pattern* is clear: growth by 1. We next fix y and extract the *rows* from the table:

$y = 0$:

x	0	1	2
z	1	3	5

$y = 1$:

x	0	1	2
z	2	4	6

$y = 2$:

x	0	1	2
z	3	5	7

A *pattern* is clear: growth by 2. We have the total of six (linear) functions!

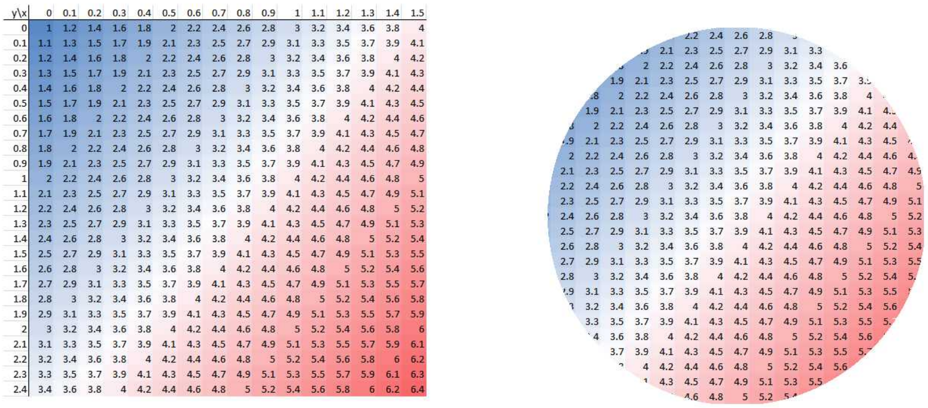
Let's do the same with a spreadsheet. This is the data:

y\x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
0	1	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4
0.1	1.1	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1
0.2	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2
0.3	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3
0.4	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4
0.5	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5
0.6	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6
0.7	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7
0.8	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8
0.9	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9
1	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5
1.1	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1
1.2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2
1.3	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3
1.4	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4
1.5	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5
1.6	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6
1.7	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7
1.8	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8
1.9	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9
2	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6
2.1	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9	6.1
2.2	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2
2.3	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9	6.1	6.3
2.4	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2	6.4

The value in each cell is computed from the corresponding value of x (all the way up) and from the corresponding value of y (all the way left). This is the formula:

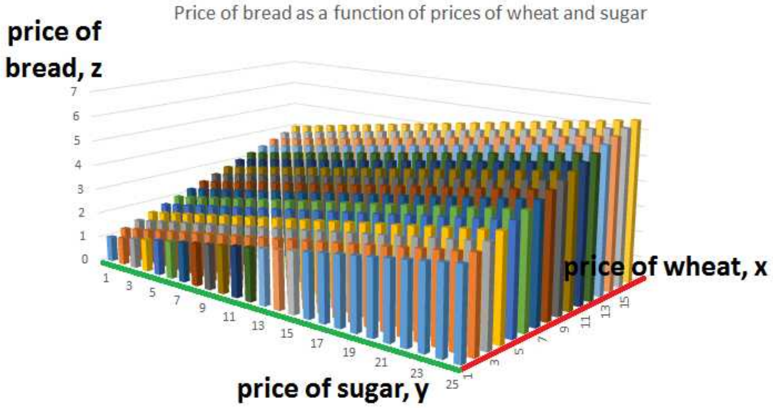
```
=2*R3C+RC2+1
```

The simplest way to visualize is by coloring the cell depending on the values (common in cartography: elevation, temperature, humidity, precipitation, population density, etc.:

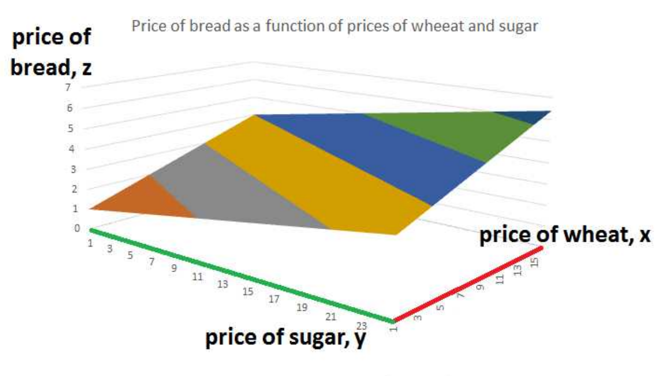


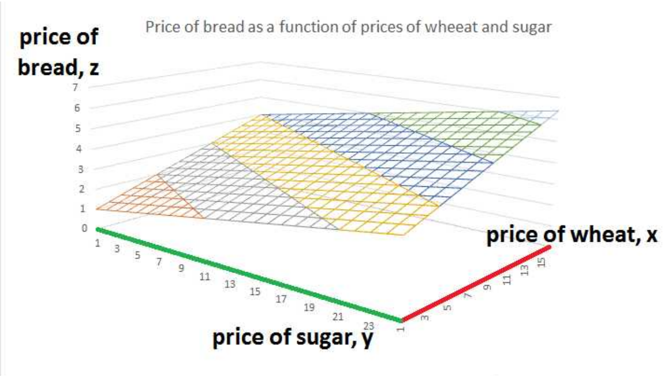
The growth is visible: It grows the most in some diagonal direction but it's *not* 45 degrees.

We can also visualize with a bar chart, just as before:



If we used bars to represent the Riemann sums to compute the *area*, here we are after the *volume*... The most common way, however, to visualize a function of two variables in mathematics is with its *graph*, which, in this case, is a surface:





In this particular case, this is a *plane*. The second graph is the same surface but displayed as a wire-frame (or even a wire-fence). The wires are the graphs of those linear functions of one variable created from our function when we fix one variable at a time. Each of these wires comes from choosing either:

- the row of x's (top) and one other row in the table, or
- the column of y's (leftmost) and one other column in the table.

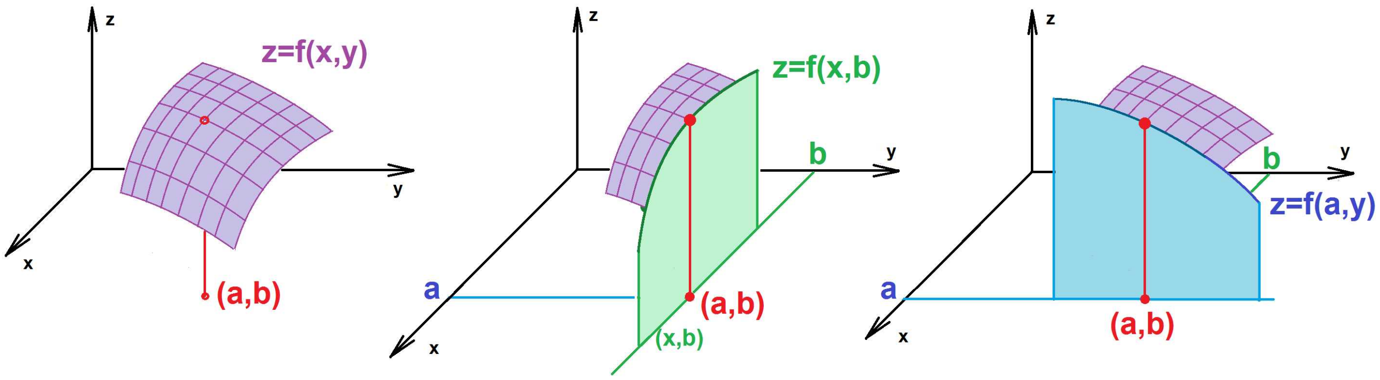
Exercise 4.3.8

Provide a similar analysis for (a) the wind-chill and (b) the heat index.

The functions of one variable created from our function $z = p(x, y)$ when we fix one variable at a time are:

$$\begin{array}{lll} y = b & \longrightarrow & f_b(x) = p(x, b) \\ x = a & \longrightarrow & g_a(y) = p(a, y) \end{array}$$

There are infinitely many of them. Their graphs are the slices – along the axes – of the surface that is the graph of F.



Therefore, the monotonicity of these functions tells us about the monotonicity of p – in the directions of the axes!

Functions of two variables are functions.

This idea comes with certain questions to be answered. What is the input, the *independent variable*? Taking a clue from our analysis of parametric curves, we answer: It is the "combination" of the two inputs of the function, i.e., x and y that form a pair, $X = (x, y)$, which is a point on the xy-plane. What is the output, the *dependent variable*? It is z.

We represent a function p diagrammatically as a *black box* that processes the input and produces the output:

$$\begin{array}{ccccc} \text{inputs} & & \text{function} & & \text{output} \\ x & & & & \\ & \searrow & & & \\ & & \boxed{\;p\;} & \mapsto & z \\ & \nearrow & & & \\ y & & & & \end{array}$$

Instead, we would like to see a single input variable, (x, y), decomposed into two, x and y, to be processed by the function *at the same time*:

$$(x, y) \;\to\; \boxed{\;p\;} \;\to\; z$$

The difference from all the functions we have seen so far is the nature of the input.

Next, what is the *domain* of p? It is supposed to be a recording of all possible inputs, i.e., all pairs (x, y) for which the output $z = p(x, y)$ of the function makes sense. This requirement creates a subset of the xy-plane and, therefore, a relation between x and y.

What about the image, i.e., the range of values of p? It is a recording of all possible outputs of p.

Definition 4.3.9: image of function of two variables

The *image* of a function of two variables $z = p(x, y)$ is the set of all such values z on the z-axis.

What about the graph of $p = (f, g)$? It is supposed to be a recording of all possible combinations of inputs and outputs of F.

Definition 4.3.10: graph of function of two variables

The *graph* of a function of two variables $z = p(x, y)$ is the set of all such points $\big(x, y, p(x, y)\big)$ in the xyz-space.

Exercise 4.3.11

Sketch the parametric curve:

$$x = t^2 - 1, \; y = 2t^2 + 3\,.$$

Exercise 4.3.12

Suppose that the following parametric curve represents the motion of an object on the plane:

$$x = 3t - 1, \; y = t^2 - 1\,.$$

(a) When does the object cross the x-axis? (b) When does the object cross the y-axis?

Exercise 4.3.13

Represent as a parametric curve the rotation of a rod of length 2 that makes one full turn every 3 seconds.

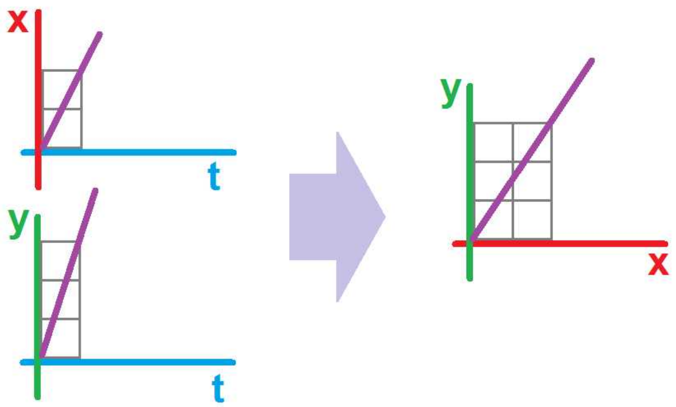
4.4. Introduction to calculus of several variables

When a parametric curve is formed from two functions of one variable:

$$x = f(t),\ y = g(t)\,,$$

are the derivatives of f and g visible in the shape (and the slope) of the path? Conversely, can the derivatives be deduced from the slopes or directions of the path?

The slopes of the graphs of f and g produce the slope of the parametric curve according to a simple rule which is easy to discover from the case when both functions are linear:



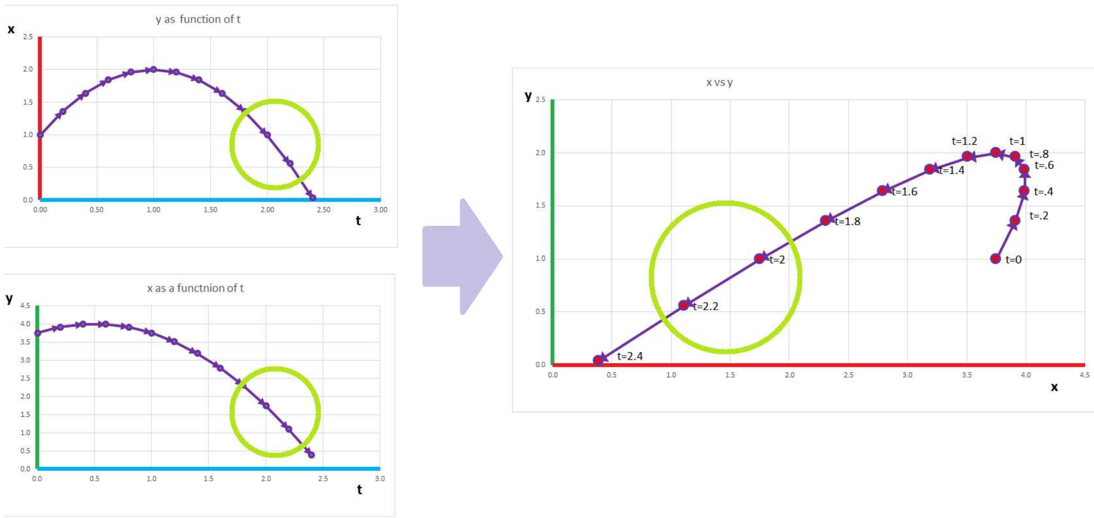
In other words, if m and n are the slopes of f and g respectively, then the slope of (f, g) is $\frac{n}{m}$. Indeed,

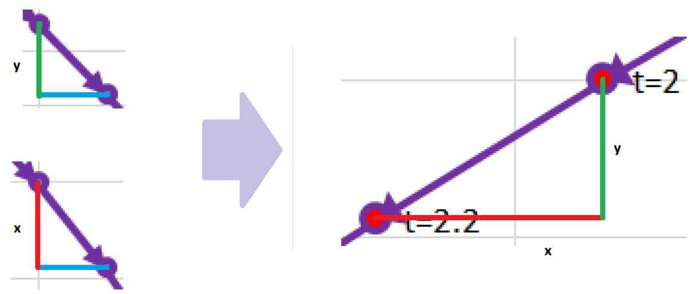
$$\frac{\text{change of } y}{\text{change of } x} = \frac{\text{change of } y/\text{change of } t}{\text{change of } x/\text{change of } t}\,.$$

When the functions are non-linear, the rule is the same but it is applied one point at a time:

$$\text{slope at } (a, b) \;=\; \frac{g'(b)}{f'(a)}\,.$$

To see why, it suffices to zoom in one a point of the parametric curve as well as the corresponding points of the graphs of the two functions:





In other words, we have this:

$$\boxed{\frac{dy}{dx} = \frac{dy}{dt} \Big/ \frac{dx}{dt}}$$

The formula resembles the *Chain Rule*, not by coincidence.

Exercise 4.4.1

Prove the formula for the case when f is one-to-one.

The two special cases are the following:

- $f'(a) = 0 \implies$ the slope is vertical, and
- $g'(b) = 0 \implies$ the slope is horizontal.

The former case was seen as "extreme" in calculus of one variable. It's not extreme in calculus of parametric curves!

Example 4.4.2: commoodities trader

Recall that the price of wheat and the price of sugar are represented by two functions combined into a parametric curve. With these functions sampled, we compute their rates of change:

$$\begin{array}{c|cc}
t & x & x' \\
\hline
0 & 1.00 & \\
\downarrow & \downarrow & \\
1 & 0.50 & \dfrac{0.5-1.0}{1-0} = -.5 \\
\downarrow & \downarrow & \\
2 & 0.33 & \dfrac{0.33-0.5}{2-1} = -.17
\end{array}
\qquad
\begin{array}{c|cc}
t & y & y' \\
\hline
0 & 1.00 & \\
\downarrow & \downarrow & \\
1 & 2.00 & \dfrac{2.00-1.00}{1-0} = 1 \\
\downarrow & \downarrow & \\
2 & 1.00 & \dfrac{1.00-2.00}{2-1} = -1
\end{array}$$

Note that there are fewer numbers than in the original because there are fewer segments than points.

Let's now confirm this result via actual differentiation of our functions:

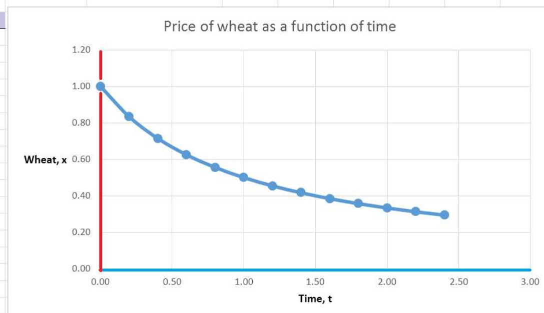
$$\begin{aligned} x &= f(t) &&= \frac{1}{t+1} \\ y &= g(t) &&= -(x-1)^2+2 \end{aligned} \quad \Longrightarrow \quad \begin{aligned} x' &= f'(t) &&= -\frac{1}{(t+1)^2} \\ y' &= g'(t) &&= -2(x-1) \end{aligned}$$

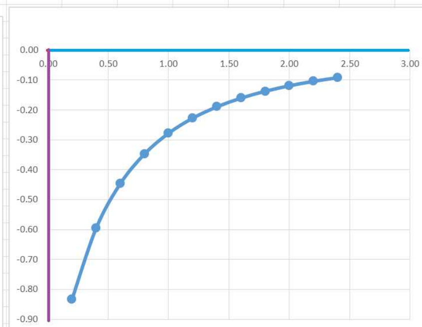
We have computed the derivatives of the two component functions. Combined, they also form a parametric curve!

The signs of the two new functions tell us the increasing/decreasing behavior of the two original functions and, therefore, the direction of the parametric curve. For example, $x' < 0$ shows that the curve moves to the left and... moves down initially because $y' < 0$. The curve also moves up because $y' > 0$. Let's visualize and confirm these results with a spreadsheet (without using the computed derivatives above):

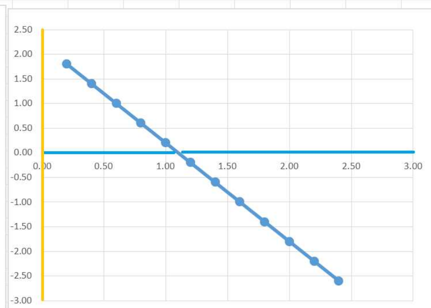
h		
0.2		

Input, time t	Output, price of wheat x	x'
0.00	1.00	
0.20	0.83	-0.83
0.40	0.71	-0.60
0.60	0.63	-0.45
0.80	0.56	-0.35
1.00	0.50	-0.28
1.20	0.45	-0.23
1.40	0.42	-0.19
1.60	0.38	-0.16
1.80	0.36	-0.14
2.00	0.33	-0.12
2.20	0.31	-0.10
2.40	0.29	-0.09





Input, time t	Output, price of sugar y	y'
0.00	1.00	
0.20	1.36	1.80
0.40	1.64	1.40
0.60	1.84	1.00
0.80	1.96	0.60
1.00	2.00	0.20
1.20	1.96	-0.20
1.40	1.84	-0.60
1.60	1.64	-1.00
1.80	1.36	-1.40
2.00	1.00	-1.80
2.20	0.56	-2.20
2.40	0.04	-2.60



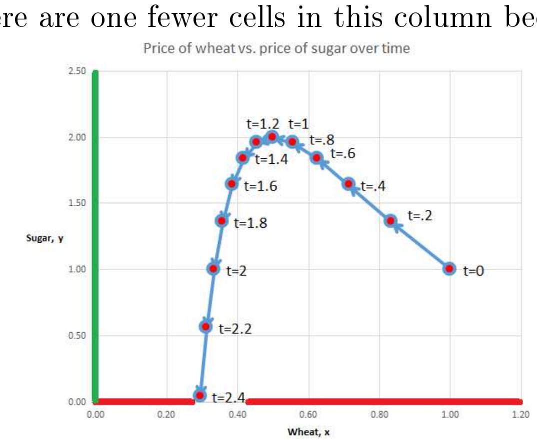
In order to approximate either of the two derivatives, we use the slope of the segment between two adjacent points. It is the *average rate of change* (also known as the difference quotient, the slope of the secant line, etc.):

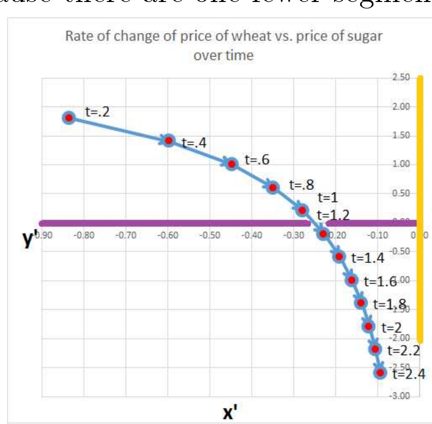
$$\frac{\text{change of } x}{\text{change of } t} \quad \text{and} \quad \frac{\text{change of } y}{\text{change of } t}.$$

The change of t is fixed as $h = \Delta t$. This value for either x and y is computed via the same spreadsheet formula as before:

```
=(RC[-1]-R[-1]C[-1])/R2C1
```

Note that there are one fewer cells in this column because there are one fewer segments than points.





In addition to $x' < 0 \implies x \searrow$, we also get $x' \nearrow \implies x \smile$ (concave up). Similarly, $y' > 0 \implies x \nearrow$ and $y' \searrow \implies y \frown$ (concave down) initially and then the opposite. Also, the apparent linearity of y'

indicate that y might be quadratic... From the monotonicity of the component functions, we conclude that initially the parametric curve goes $\nwarrow$ and then $\swarrow$ to confirm the picture.

Exercise 4.4.3

What conclusions about the shape of the parametric curve can you draw from the concavity of its component functions?

Thus, we can say about a parametric curve that *its derivative is made up of the derivatives of the functions involved.* There are only two for a parametric curve – and infinitely many for a function of two variables!

Example 4.4.4: baker

The dependence of the price of bread on the prices of wheat and sugar are represented by the function $z = p(x, y)$ below. As it is sampled, we can compute the rates of change of this function in the horizontal and vertical directions:

$$\text{over } x: \begin{array}{c|ccccc} y\backslash x & 0 & \to & 1 & \to & 2 \\ \hline 0 & 1 & \to & 3 & \to & 5 \\ \\ 1 & 2 & \to & 4 & \to & 6 \\ \\ 2 & 3 & \to & 5 & \to & 7 \end{array} \qquad \text{over } y: \begin{array}{c|ccc} y\backslash x & 0 & 1 & 2 \\ \hline 0 & 1 & 3 & 5 \\ \downarrow & \downarrow & \downarrow & \downarrow \\ 1 & 2 & 4 & 6 \\ \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 3 & 5 & 7 \end{array}$$

Recall that by fixing one of the variables, we create a function of one variable with respect to the other variable. Now we approximate the derivatives of these functions just as before, via the *average rate of change*:

$$\frac{\text{change of } z}{\text{change of } x} \text{ and } \frac{\text{change of } z}{\text{change of } y}.$$

First, we approximate the derivative in the direction of y:

$$\begin{array}{c|cc} y\ (x=0) & z & z' \\ \hline 0\ \downarrow & 1\ \downarrow & \\ 1\ \downarrow & 2\ \downarrow & \frac{2-1}{1-0}=1 \\ 2\ \downarrow & 3\ \downarrow & \frac{3-2}{2-1}=1 \end{array} \qquad \begin{array}{c|cc} y\ (x=1) & z & z' \\ \hline 0\ \downarrow & 3\ \downarrow & \\ 1\ \downarrow & 4\ \downarrow & \frac{4-3}{1-0}=1 \\ 2\ \downarrow & 5\ \downarrow & \frac{5-4}{2-1}=1 \end{array} \qquad \begin{array}{c|cc} y\ (x=2) & z & z' \\ \hline 0\ \downarrow & 5\ \downarrow & \\ 1\ \downarrow & 6\ \downarrow & \frac{6-5}{2-1}=1 \\ 2\ \downarrow & 7\ \downarrow & \frac{7-6}{2-1}=1 \end{array}$$

All 1s. Note that there are fewer numbers than in the original because there are fewer segments than points. Similarly, we approximate the derivative in the direction of x:

$$\begin{array}{r|cccc} (y=0)\ x & 0 & \to 1 & & \to 2 \\ \hline z & 1 & \to 3 & & \to 5 \\ \hline z' & & \frac{3-1}{1-0}=2 & & \frac{5-3}{2-1}=2 \end{array} \qquad \begin{array}{r|ccc} (y=1)\ x & 0 & 1 & 2 \\ \hline z & 2 & 4 & 6 \\ \hline z' & & 2 & 2 \end{array} \qquad \begin{array}{r|ccc} (y=0)\ x & 0 & 1 & 2 \\ \hline z & 3 & 5 & 7 \\ \hline z' & & 2 & 2 \end{array}$$

All 2s. We put these one-variable functions together; then the rates of change of F with respect to x and y are these new functions of two variables respectively:

$$\text{leading to } \begin{array}{c|ccc} y\backslash x & 0 & 1 & 2 \\ \hline 0 & & 2 & 2 \\ 1 & & 2 & 2 \\ 2 & & 2 & 2 \end{array} \ \& \ \begin{array}{c|ccc} y\backslash x & 0 & 1 & 2 \\ \hline 0 & & & \\ 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 1 \end{array} \text{ leading to } \begin{array}{c|ccc} y\backslash x & 0 & 1 & 2 \\ \hline 0 & & & \\ 1 & & (2,1) & (2,1) \\ 2 & & (2,1) & (2,1) \end{array}$$

The two functions are further combined on the right. As we shall see later, going 2 horizontally and 1 vertically is the direction of the fastest growth of the function p.

Let's now confirm this result via actual differentiation of our function:

$$p(x,y)=2x+y+1\,.$$

Just as before, we fix one of the variables and differentiate with respect to the other. We call these two functions *the partial derivatives of* p *with respect to* x *and* y respectively.

We use the following two types of notation (following Leibniz and Lagrange as before). For x, we declare y fixed and differentiate over x

$$\frac{\partial p}{\partial x}=p'_x=\frac{\partial}{\partial x}\,(2x+y+1)=\frac{\partial}{\partial x}(2x)+0+0=2\,.$$

For y, we declare x fixed and differentiate over y:

$$\frac{\partial p}{\partial y}=p'_y=\frac{\partial}{\partial y}\,(2x+y+1)=0+\frac{\partial}{\partial y}(y)+0=1\,.$$

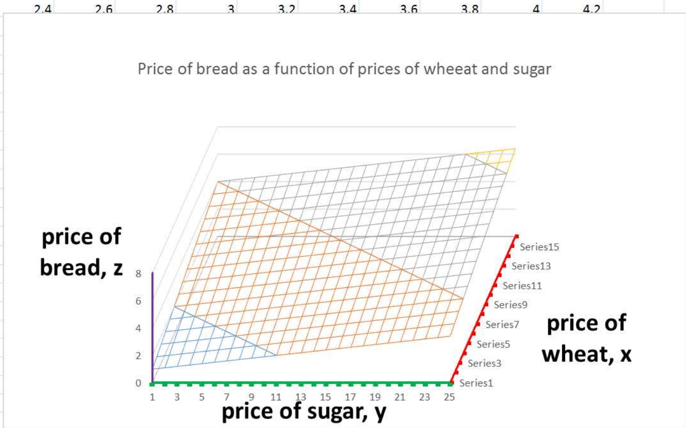
The conclusion might sound familiar: The derivative of a linear function is constant! The combination of the two partial derivatives will be seen as the derivative of p called the *gradient* of p. This is a new kind of function to be discussed in Volume 4 (Chapters 4HD-4 and 4HD-5).

We can confirm these results by examining the spreadsheet. Each line (wire) below on the right is the graph of a function of one variable:

h

0.1

y\x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
0	1	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4
0.1	1.1	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1
0.2	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2
0.3	1.3	1.5	1.7	1.9	2.1	2.3										
0.4	1.4	1.6	1.8	2	2.2	2.4										
0.5	1.5	1.7	1.9	2.1	2.3	2.5										
0.6	1.6	1.8	2	2.2	2.4	2.6										
0.7	1.7	1.9	2.1	2.3	2.5	2.7										
0.8	1.8	2	2.2	2.4	2.6	2.8										
0.9	1.9	2.1	2.3	2.5	2.7	2.9										
1	2	2.2	2.4	2.6	2.8	3										
1.1	2.1	2.3	2.5	2.7	2.9	3.1										
1.2	2.2	2.4	2.6	2.8	3	3.2										
1.3	2.3	2.5	2.7	2.9	3.1	3.3										
1.4	2.4	2.6	2.8	3	3.2	3.4										
1.5	2.5	2.7	2.9	3.1	3.3	3.5										
1.6	2.6	2.8	3	3.2	3.4	3.6										
1.7	2.7	2.9	3.1	3.3	3.5	3.7										
1.8	2.8	3	3.2	3.4	3.6	3.8										
1.9	2.9	3.1	3.3	3.5	3.7	3.9										
2	3	3.2	3.4	3.6	3.8	4										
2.1	3.1	3.3	3.5	3.7	3.9	4.1										
2.2	3.2	3.4	3.6	3.8	4	4.2										
2.3	3.3	3.5	3.7	3.9	4.1	4.3										
2.4	3.4	3.6	3.8	4	4.2	4.4										



And each has its own derivative! As we move horizontally, the values of x grow by 0.1 while the values of z grow by 0.2. Therefore, $p'_x=2$. Similarly, as we move vertically, the values of y grow by 0.1 and so do the values of z. Therefore, $p'_y=1$.

Warning!

Do not confuse partial differentiation with implicit differentiation that comes under *related rates*:

$$\frac{\partial}{\partial x}\,(xy)=y\quad\text{vs.}\quad\frac{d}{dx}\,(xy)=y+\frac{dy}{dx}\,.$$

The extra term on the right comes from the fact that the two variables are *related*.

Example 4.4.5: non-linear

Let's consider this function again:

$$q(x,y)=\sin(xy)\,.$$

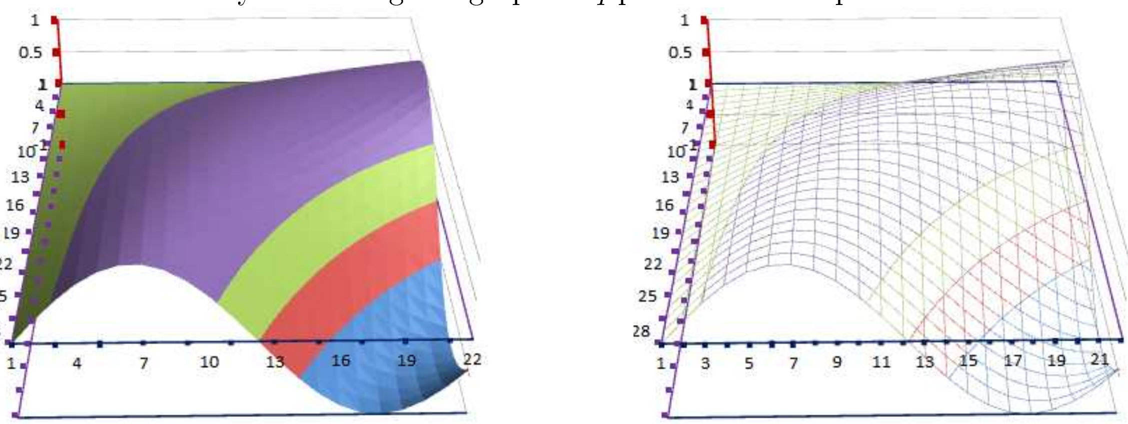
Compute the partial derivatives by the *Chain Rule*. First we declare y an unknown and unspecified but fixed *parameter* and carry out differentiation with respect to x:

$$\frac{\partial q}{\partial x}=\frac{\partial}{\partial x}\big(\sin(xy)\big)=\cos(xy)\cdot\frac{\partial}{\partial x}(xy)=\cos(xy)y\,.$$

This time, x is the parameter:

$$\frac{\partial q}{\partial y}=\frac{\partial}{\partial y}\big(\sin(xy)\big)=\cos(xy)\cdot\frac{\partial}{\partial y}(xy)=\cos(xy)x\,.$$

Let's confirm these results by examining the graph of q plotted with a spreadsheet:



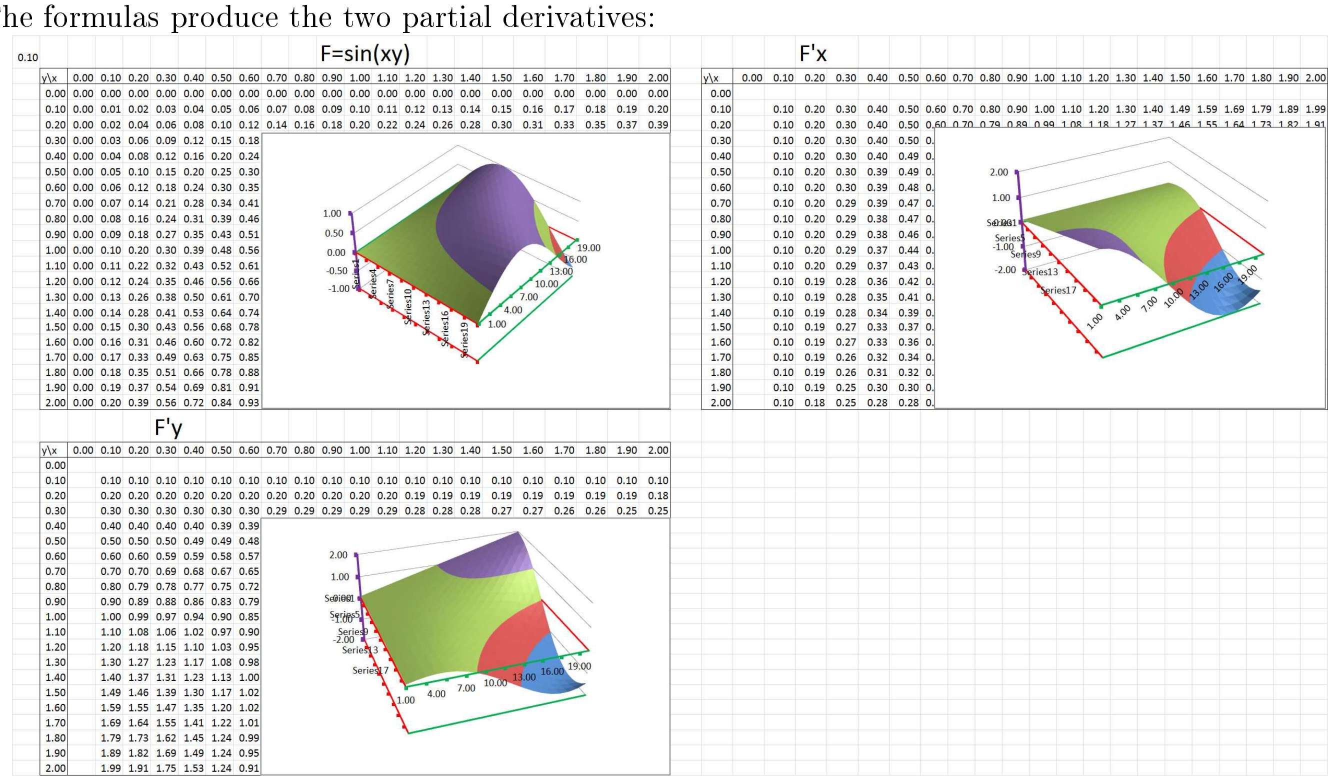
Note that the edge of the surface is a curve and it is the graph of the function given in the very last row of the table. We also notice that:

- The surface is flat along the x-axis, because $\frac{\partial q}{\partial x}=0$ for $y=0$.
- The surface is flat along the y-axis, because $\frac{\partial q}{\partial x}=0$ for $x=0$.

We approximate these derivatives just as before:The average rate of change with the change of x and y is fixed as $h=\Delta x=\Delta y$. In a spreadsheet, the rate of change of z in the direction of x and y is computed via the same formula but applied horizontally and vertically respectively:

`=(R[-23]C-R[-23]C[-1])/R1C1` and `=(RC[-23]-R[-1]C[-23])/R1C1`

The formulas produce the two partial derivatives:



0.10

F=sin(xy)

y\x	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.40	1.50	1.60	1.70	1.80	1.90	2.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.10	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20
0.20	0.00	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20	0.22	0.24	0.26	0.28	0.30	0.31	0.33	0.35	0.37	0.39
0.30	0.00	0.03	0.06	0.09	0.12	0.15	0.18														
0.40	0.00	0.04	0.08	0.12	0.16	0.20	0.24														
0.50	0.00	0.05	0.10	0.15	0.20	0.25	0.30														
0.60	0.00	0.06	0.12	0.18	0.24	0.30	0.35														
0.70	0.00	0.07	0.14	0.21	0.28	0.34	0.41														
0.80	0.00	0.08	0.16	0.24	0.31	0.39	0.46														
0.90	0.00	0.09	0.18	0.27	0.35	0.43	0.51														
1.00	0.00	0.10	0.20	0.30	0.39	0.48	0.56														
1.10	0.00	0.11	0.22	0.32	0.43	0.52	0.61														
1.20	0.00	0.12	0.24	0.35	0.46	0.56	0.66														
1.30	0.00	0.13	0.26	0.38	0.50	0.61	0.70														
1.40	0.00	0.14	0.28	0.41	0.53	0.64	0.74														
1.50	0.00	0.15	0.30	0.43	0.56	0.68	0.78														
1.60	0.00	0.16	0.31	0.46	0.60	0.72	0.82														
1.70	0.00	0.17	0.33	0.49	0.63	0.75	0.85														
1.80	0.00	0.18	0.35	0.51	0.66	0.78	0.88														
1.90	0.00	0.19	0.37	0.54	0.69	0.81	0.91														
2.00	0.00	0.20	0.39	0.56	0.72	0.84	0.93														

F'x

y\x	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.40	1.50	1.60	1.70	1.80	1.90	2.00
0.00																					
0.10		0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.40	1.49	1.59	1.69	1.79	1.89	1.99
0.20		0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.79	0.89	0.99	1.08	1.18	1.27	1.37	1.46	1.55	1.64	1.73	1.82	1.91
0.30		0.10	0.20	0.30	0.40	0.50	0.														
0.40		0.10	0.20	0.30	0.40	0.49	0.														
0.50		0.10	0.20	0.30	0.39	0.49	0.														
0.60		0.10	0.20	0.30	0.39	0.48	0.														
0.70		0.10	0.20	0.29	0.39	0.47	0.														
0.80		0.10	0.20	0.29	0.38	0.47	0.														
0.90		0.10	0.20	0.29	0.38	0.46	0.														
1.00		0.10	0.20	0.29	0.37	0.44	0.														
1.10		0.10	0.20	0.29	0.37	0.43	0.														
1.20		0.10	0.19	0.28	0.36	0.42	0.														
1.30		0.10	0.19	0.28	0.35	0.41	0.														
1.40		0.10	0.19	0.28	0.34	0.39	0.														
1.50		0.10	0.19	0.27	0.33	0.37	0.														
1.60		0.10	0.19	0.27	0.33	0.36	0.														
1.70		0.10	0.19	0.26	0.32	0.34	0.														
1.80		0.10	0.19	0.26	0.31	0.32	0.														
1.90		0.10	0.19	0.25	0.30	0.30	0.														
2.00		0.10	0.18	0.25	0.28	0.28	0.														

F'y

y\x	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.40	1.50	1.60	1.70	1.80	1.90	2.00
0.00																					
0.10		0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
0.20		0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.18
0.30		0.30	0.30	0.30	0.30	0.30	0.30	0.29	0.29	0.29	0.29	0.29	0.28	0.28	0.28	0.27	0.27	0.26	0.26	0.25	0.25
0.40		0.40	0.40	0.40	0.40	0.39	0.39														
0.50		0.50	0.50	0.50	0.49	0.49	0.48														
0.60		0.60	0.60	0.59	0.59	0.58	0.57														
0.70		0.70	0.70	0.69	0.68	0.67	0.65														
0.80		0.80	0.79	0.78	0.77	0.75	0.72														
0.90		0.90	0.89	0.88	0.86	0.83	0.79														
1.00		1.00	0.99	0.97	0.94	0.90	0.85														
1.10		1.10	1.08	1.06	1.02	0.97	0.90														
1.20		1.20	1.18	1.15	1.10	1.03	0.95														
1.30		1.30	1.27	1.23	1.17	1.08	0.98														
1.40		1.40	1.37	1.31	1.23	1.13	1.00														
1.50		1.49	1.46	1.39	1.30	1.17	1.02														
1.60		1.59	1.55	1.47	1.35	1.20	1.02														
1.70		1.69	1.64	1.55	1.41	1.22	1.01														
1.80		1.79	1.73	1.62	1.45	1.24	0.99														
1.90		1.89	1.82	1.69	1.49	1.24	0.95														
2.00		1.99	1.91	1.75	1.53	1.24	0.91														

At the point $(0.1, 0.1)$, the values of the two partial derivatives are equal, which is why the direction of the fastest growth of q is at 45 degrees. Note also that the highest locations form a ridge; it is where both partial derivatives are equal to 0. To find the location of the ridge, we solve the equation

$$\cos(xy) = 0 \implies xy = \pi/2\,.$$

It's a hyperbola.

The idea that the values of the partial derivatives indicate the direction of the fastest growth of the function can be illustrated approximately as follows:

$y \quad \backslash x$	$q'_x > 0$	$q'_x < 0$
$q'_y > 0$	$\nearrow$	$\nwarrow$
$q'_y < 0$	$\searrow$	$\swarrow$

Example 4.4.6: bread buyers

We will take the two examples – the commodity trader and the baker – from the last two sections and ask, what price of bread have daily visitors to the bakery shop seen over time? These are the variables:

- t is time.

Two variables representing these two *commodities*:

- x is the price of wheat.
- y is the price of sugar.

And a *product*:

- z is the price of a loaf of bread.

The visitors see how z depends on t, via some function of *single variable*:

$$z = h(t)\,.$$

What is it?

This is the summary of the setup:

Example 1 (trader)				
t	$\longrightarrow$	(x, y)	$\longrightarrow$	z
	Example 2 (baker)			

We realize that the problem is about *compositions*!

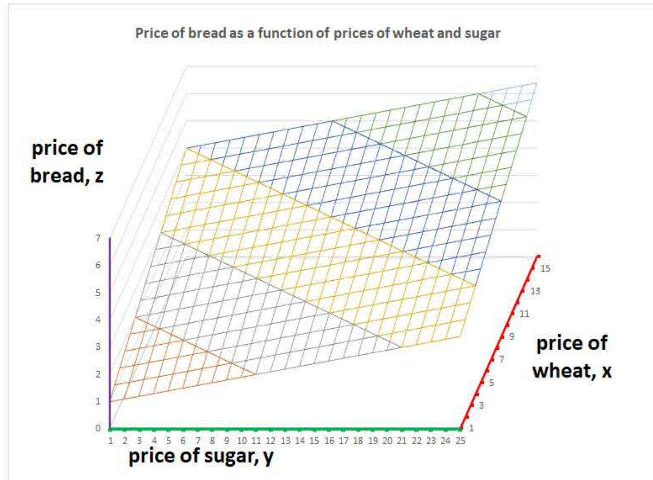
Recall that the price of wheat and the price of sugar are represented by a parametric curve:

$$x = f(t) = \frac{1}{t+1},\ y = g(t) = -(x-1)^2 + 2\,.$$

Furthermore, the price of bread is computed from the other two prices by the function of two variables:

$$z = p(x, y) = 2x + y + 1\,.$$

The two functions are visualized as follows:



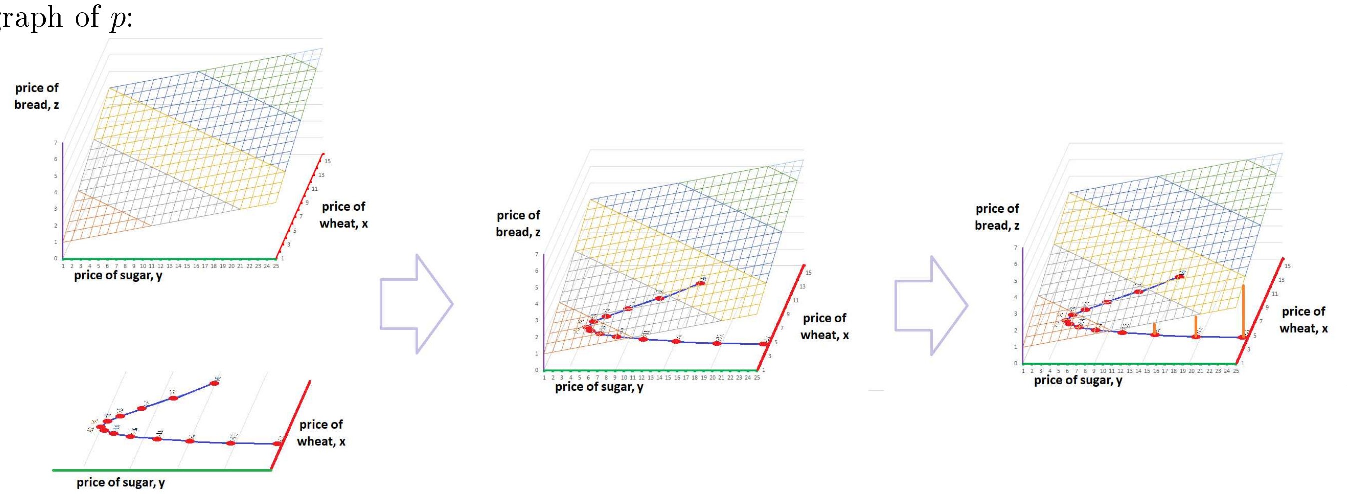
Then, of course, h is the composition of these two:

$$t \mapsto (x, y) \mapsto z\,,$$

computed via the following substitution:

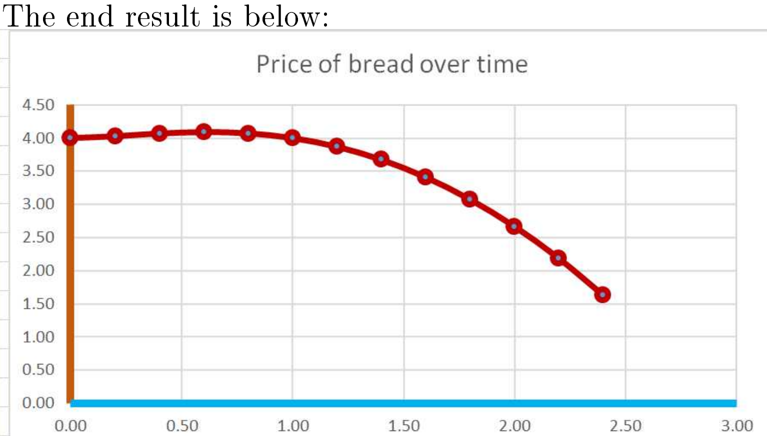
$$h(t) = p\big(f(t), g(t)\big)\,.$$

To visualize what happens, imagine the parametric curve – on the xy-plane – being "lifted" to the graph of p:



The elevation is then the value of h. The end result is below:

t	x	y	z
0.00	1.00	1.00	4.00
0.20	0.83	1.36	4.03
0.40	0.71	1.64	4.07
0.60	0.63	1.84	4.09
0.80	0.56	1.96	4.07
1.00	0.50	2.00	4.00
1.20	0.45	1.96	3.87
1.40	0.42	1.84	3.67
1.60	0.38	1.64	3.41
1.80	0.36	1.36	3.07
2.00	0.33	1.00	2.67
2.20	0.31	0.56	2.19
2.40	0.29	0.04	1.63



In the past, we have found the derivative of the composition of two functions by the *Chain Rule*: We expressed it in terms of the derivatives of the two functions involved (their product). We then conjecture that in order to find the derivative of the composition above, we need to understand the meaning of the following:

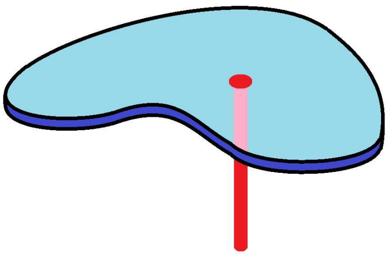
- the derivative of a parametric curve, and

- the derivative of a function of two variables.

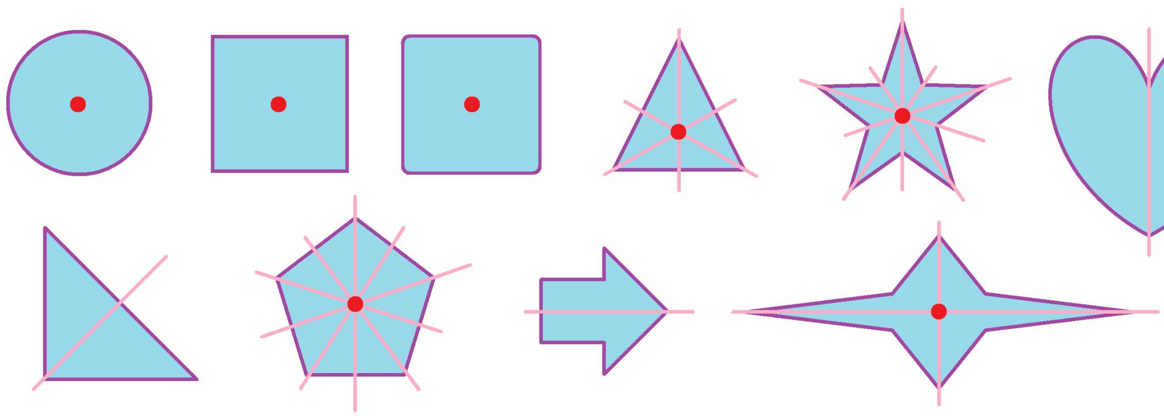
4.5. The centroid of a flat object

We continue with our analysis of the center of mass started in Chapter 3 but in the 2-dimensional setting.

Suppose we have a plate with uniform a density and identical thickness (it is known as a “lamina”). How can we balance it on a single support called the *centroid*?

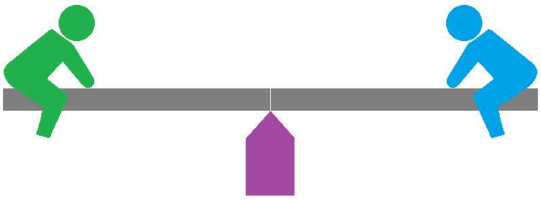


There are a few heuristics that help. If the object has a “center”, such as a circle or a square, this is it.



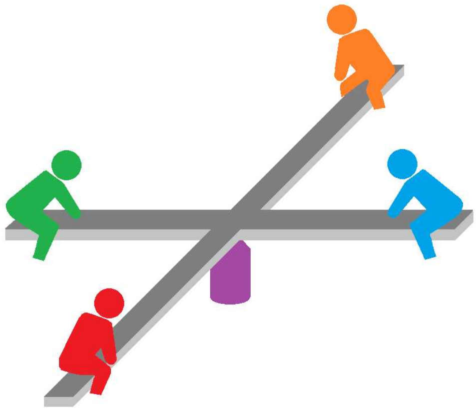
Also, any axis of symmetry will have to contain the centroid.

The idea of centroid is related to the concept of the *center of mass* which is the center of rotation of the object when subjected to a force. We studied this concept previously but with the weight distributed within a straight segment, such as a seesaw:



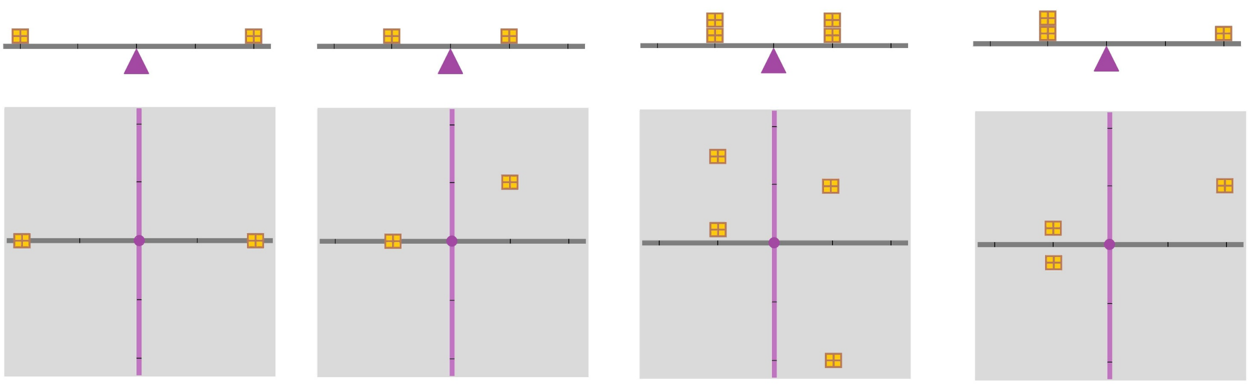
We found that if one person is heavier than the other, the latter person should sit farther from the center in order to balance the beam. In fact, the distance should be twice as long!

Consider a variation of the *seesaw*. It is made of two beams nailed together to form a cross with a single point of support in the middle:

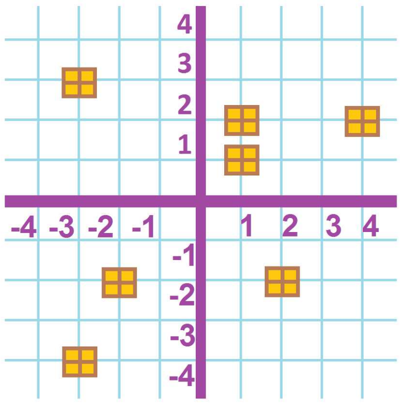


It appears that four persons of equal weight will be in balance when located at equal distance from the point of support. But there is more: They will be balanced as long as either *pair* of persons facing each other are in balance! We can then use what we have learned from the 1-dimensional case.

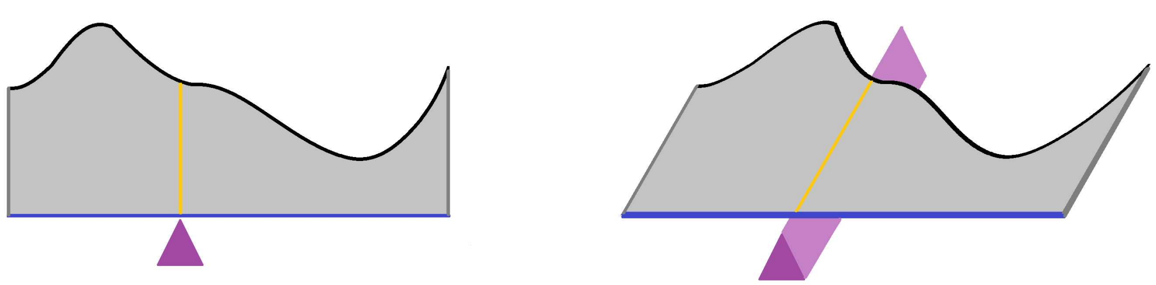
We explore this idea by replacing this construction with a square. Then the seesaws that we considered previously can be interpreted as this square *balanced on a bar* that goes all the way across:



We can spread the weight along the line parallel to the bar because only the distance to this bar matters for the leverage of each weight. Once we add the x- and y-axis to the picture, this distance is simply the x-coordinate:



Now, our problem is that of balancing the region below the graph of a function.



Let's review how we do this. Suppose we have a non-negative function $y = f(x)$ integrable on segment $[a, b]$. For a given point c, the integral

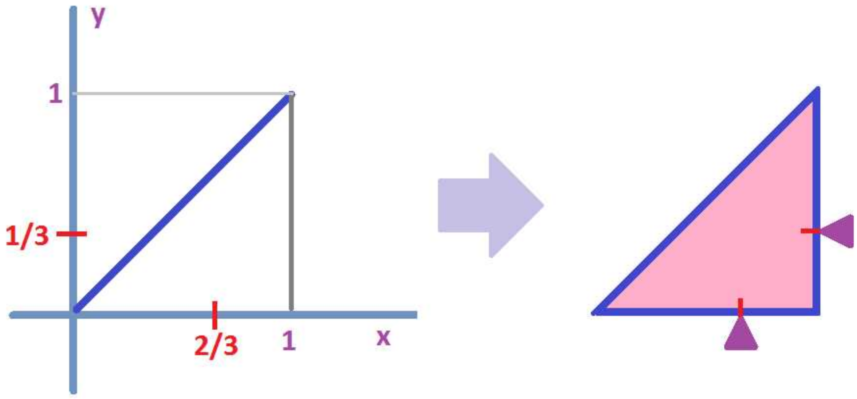
$$\int_a^b f(x)(x-c)\,dx$$

is called the *total moment of the region with respect to* c. The *center of mass of the region* is such a point c that the total moment with respect to c is zero:

$$c = \frac{\int_a^b f(x)x\,dx}{\int_a^b f(x)\,dx}.$$

Example 4.5.1: balance a triangle

Let's review how we can balance a triangle on its horizontal edge. Suppose it is the region under the graph $y = f(x) = x$ from 0 to 1:



We compute the total moment of the object:

$$\begin{aligned}\int_0^1 f(x)x\,dx &= \int_0^1 x\cdot x\,dx\\ &= \int_0^1 x^2\,dx\\ &= x^3/3\Big|_0^1\\ &= 1/3\,.\end{aligned}$$

Meanwhile, the mass is simply $1/2$. Therefore, the center of mass – on the x-axis – is

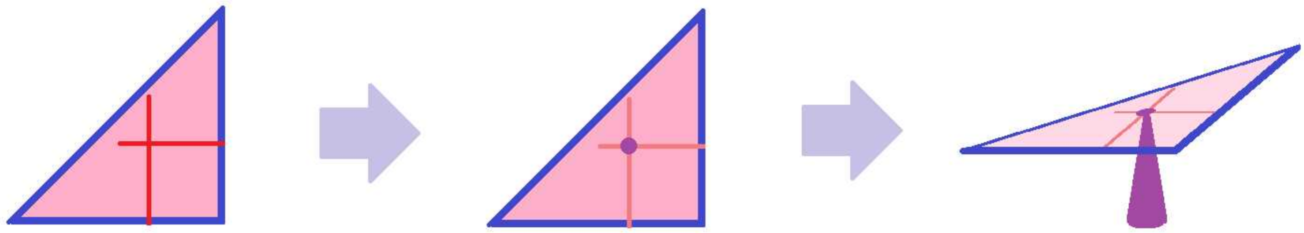
$$c_1 = \frac{1}{3} \div \frac{1}{2} = \frac{2}{3}\,.$$

What if we want to balance the triangle on its other edge? We place the x-axis along that edge, then the slanted edges is given by $y = g(x) = 1 - x$. We compute the total moment of the object:

$$\begin{aligned}\int_0^1 g(x)x\,dx &= \int_0^1 (1-x)x\,dx\\ &= \int_0^1 (x-x^2)\,dx\\ &= y^2/2 - y^3/3\Big|_0^1\\ &= 1/2 - 1/3\\ &= 1/6\,.\end{aligned}$$

Meanwhile, the mass is still 1/2. Therefore, the center of mass – along this edge – is

$$c_2 = \frac{1}{6} \div \frac{1}{2} = \frac{1}{3}\,.$$



We can balance the triangle on either of two bars. Now we remove the bars and replace them with a single support placed at their intersection.

Definition 4.5.2: total moment

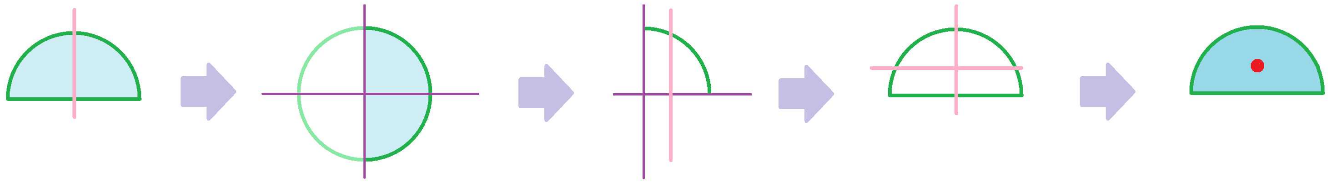
Suppose a function $y = f(x)$ is integrable on $[a, b]$. Then the *total moment* of the region under the graph of f with respect to the line $x = c$ is defined to be:

$$\boxed{\int_a^b (x - c) f(x)\, dx}$$

Such a line is an *axis* of the region if the total moment is zero.

Example 4.5.3: half-circle

Let's try the half-circle. One of the axes will go through the center perpendicular to the diameter. We just need to find the other. That's why we place the quarter of the disk adjacent to the origin:



Then the total moment is:

$$\int_0^1 (x - c)\sqrt{1 - x^2}\, dx = 0\,.$$

Therefore,

$$\int_0^1 x\sqrt{1 - x^2}\, dx = c\int_0^1 \sqrt{1 - x^2}\, dx\,.$$

The integral on the right-hand side is simply the area of the quarter circle, and the one on the left-hand side is easily evaluated by substitution ($u = 1 - x^2$):

$$\begin{aligned}\int_0^1 x\sqrt{1 - x^2}\, dx &= \int_1^0 -\frac{1}{2}\sqrt{u}\, du\\ &= -\frac{1}{2}\frac{2}{3}u^{3/2}\bigg|_1^0\\ &= \frac{1}{3}\,.\end{aligned}$$

Therefore, we have:

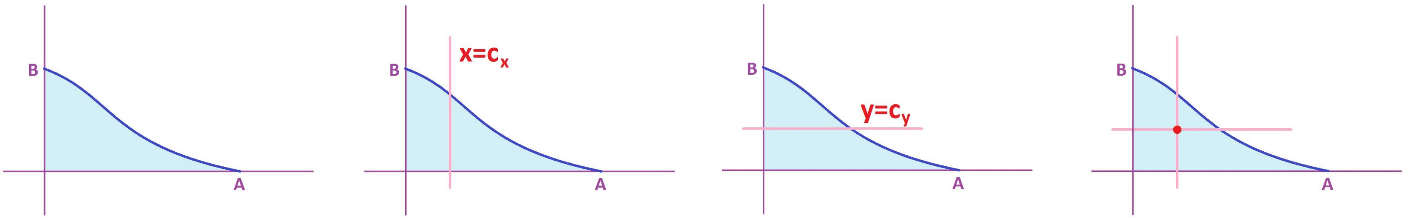
$$\frac{1}{3} = c\pi/4 \implies c = \frac{4}{3\pi} \approx .42\,.$$

Exercise 4.5.4

Find the axes of the region bounded by the graph $y = \dfrac{1}{x+0.5} - 0.5$ and the axes.

We have been able to use only this definition to find the axes of regions with symmetries.

For the general case, we will have to use the x- and y-axes available to us:



Definition 4.5.5: centroid

Suppose a function $y = f(x)$ is decreasing on $[0, A]$ and $f(0) = B > 0$. Then the *centroid* of the region bounded by the graph of f, the x-axis, and the y-axis is a point (c_x, c_y) such that the total moments of the region with respect to the lines $x = c_x$ and $y = c_y$ are zero; i.e.,

$$\int_0^A (x - c_x) f(x)\, dx = 0\,,$$

and

$$\int_0^B (y - c_y) f^{-1}(y)\, dy = 0\,.$$

Then, the coordinates of the centroid are:

$$\boxed{c_x = \frac{1}{A}\int_0^A x f(x)\, dx \quad \text{and} \quad c_y = \frac{1}{A}\int_0^B y f^{-1}(y)}$$

where A is the area of the region.

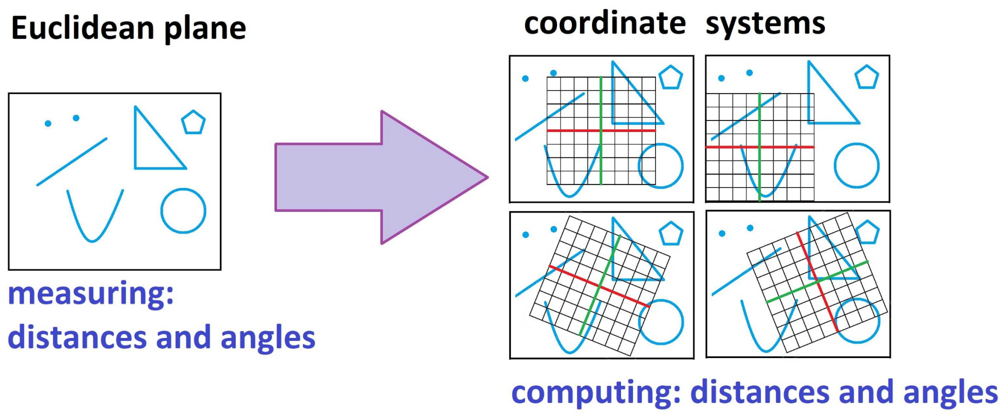
Exercise 4.5.6

Find the centroid of the region bounded by the graph of $y = 1 - x^2$.

The general case of a plate of an arbitrary shape will be addressed in Volume 4.

4.6. *Coordinate systems; polar coordinates*

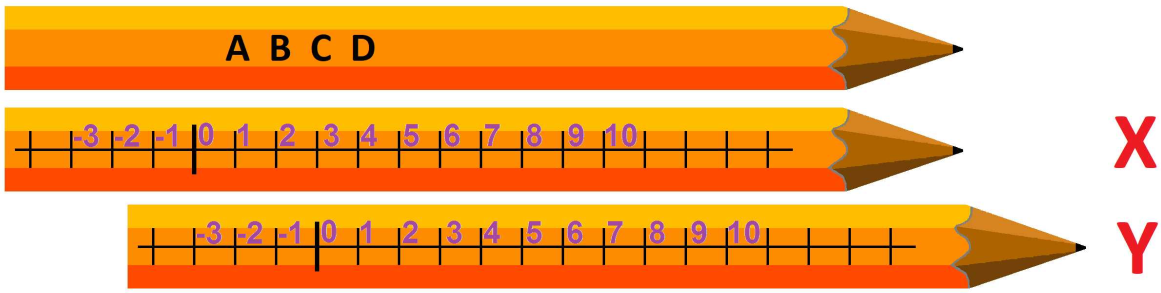
If we want to study the Euclidean plane, and Euclidean geometry, algebraically, we *superimpose* the Cartesian grid over this plane:



We can place the coordinate system on top of our physical space in a number of ways. Meanwhile the geometry on the piece of paper determines what is going on, not a particular choice of the coordinate system.

We start with dimension 1.

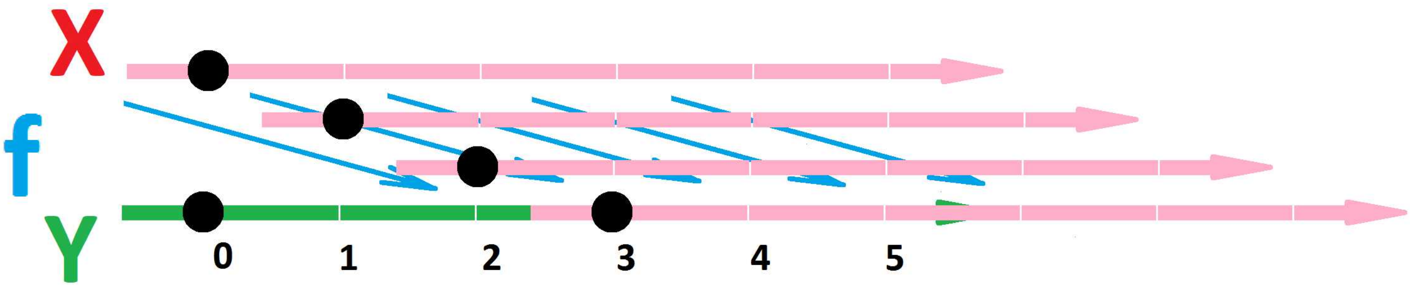
The *line* can have different coordinate systems assigned to it. We can imagine that we have three pencils: one unmarked and two with the whole x-axis is drawn on them. The first represents the "reality",and the other two represent two different coordinate systems to be used to record the locations on the first:



Then we have:

1. Point A has coordinate 1 with respect to the first Cartesian system but -2 with respect to the second.
2. Point B has coordinate 2 with respect to the first Cartesian system but -1 with respect to the second.
3. And so on.

These Cartesian systems are related to each other via some basic transformations, such as a shift:

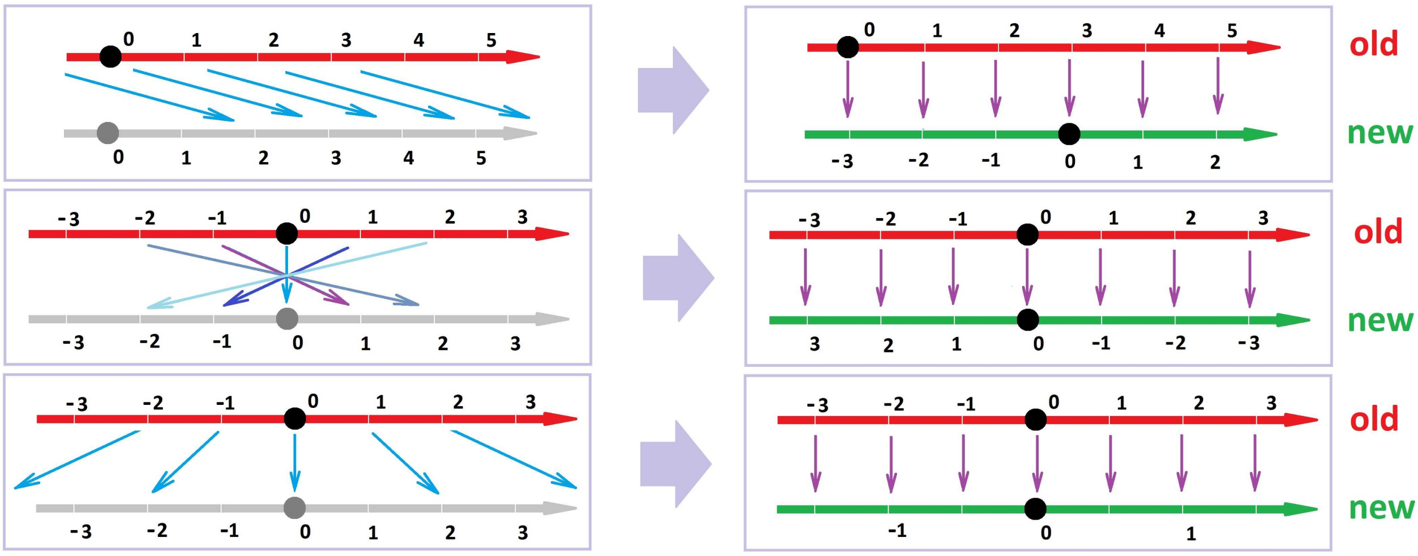


Above you see two ways to interpret the transformation:

1. The arrows are between the x-axis and the intact y-axis.
2. We move the y-axis so that $y = f(x)$ is aligned with x.

We followed the former in Chapter 1PC-2 and we will follow the latter in this section.

These are the three main transformations of an axis: shift, flip, and stretch (left). And this is what happens to the coordinates of points (right):



Warning!

The points don't move.

This is the algebra for the basic transformations of the axis, the old and the new coordinates:

$$
\begin{array}{lll}
1. \quad t & \xrightarrow{\text{shift by } k} & x = t - k \\
2. \quad t & \xrightarrow{\text{flip}} & x = -t \\
3. \quad t & \xrightarrow{\text{stretch by } k} & x = t/k
\end{array}
$$

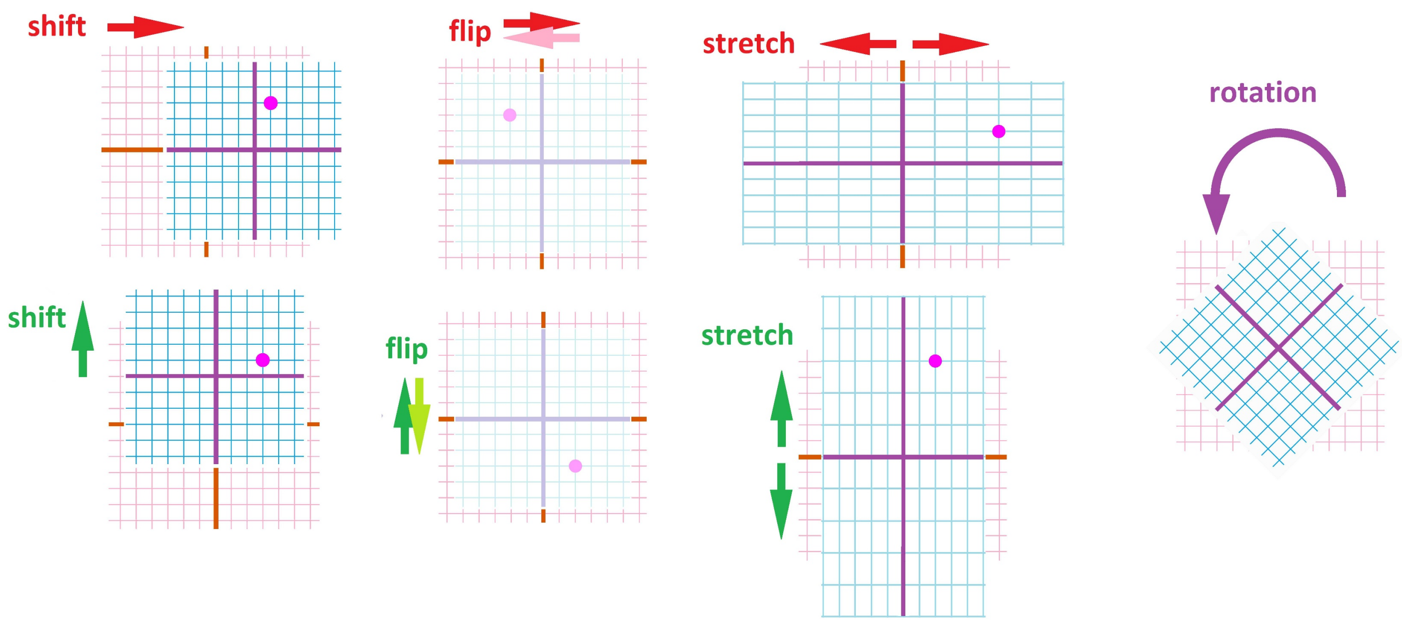
The three may come respectively from:

1. changing the starting location of the milestones
2. replacing east with west as the positive direction
3. switching from miles to kilometers

Now dimension 2, the *plane*.

Both x- and y-axis can be subjected to the transformations above. The change of coordinates under the resulting six basic transformations of the xy-plane is shown below:

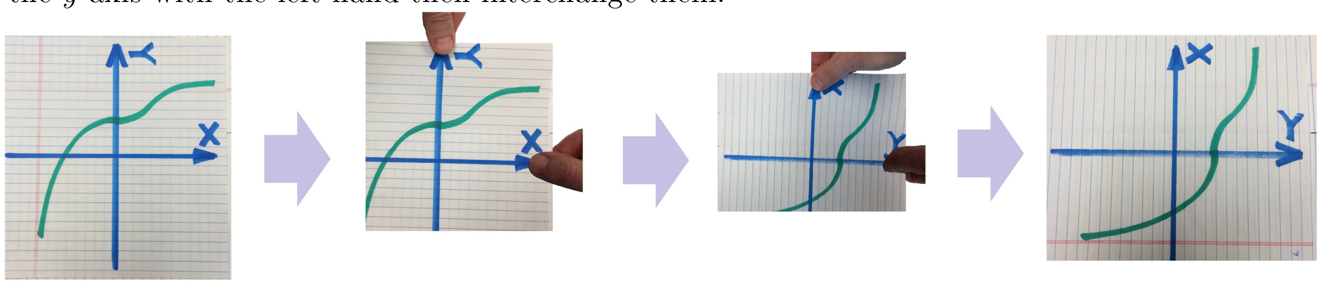
$$
\begin{array}{llll}
\text{vertical shift:} & \begin{array}{c}(\ x\ ,\ y\ )\\(\ x\ ,\ y-k\ )\end{array} & \text{flip:}\ \begin{array}{c}(\ x\ ,\ y\ )\\(\ x\ ,\ y\cdot(-1)\ )\end{array} & \text{stretch:}\ \begin{array}{c}(\ x\ ,\ y\ )\\(\ x\ ,\ y/k\ )\end{array}\\
\text{horizontal shift:} & \begin{array}{c}(\ x\ ,\ y\ )\\(\ x-k\ ,\ y\ )\end{array} & \text{flip:}\ \begin{array}{c}(\ x\ ,\ y\ )\\(\ x\cdot(-1)\ ,\ y\ )\end{array} & \text{stretch:}\ \begin{array}{c}(\ x\ ,\ y\ )\\(\ x/k\ ,\ y\ )\end{array}
\end{array}.
$$



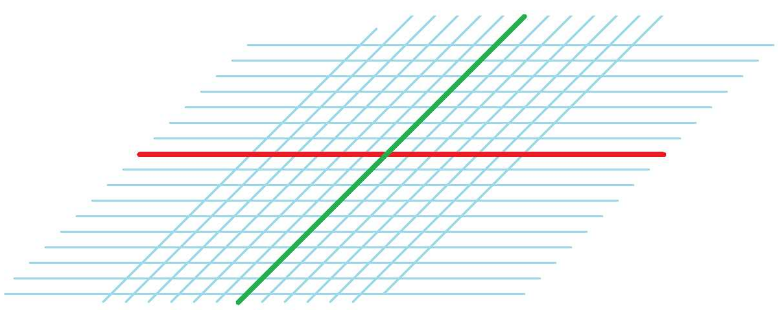
But some transformations cannot be reduced to a combination of these six, such as the rotation.

Example 4.6.1: other transformations

Recall from Chapter 1PC-3, that in order to find the graph of the inverse function, we execute a *flip about the diagonal* of the plane. We grab the end of the x-axis with the right hand and grab the end of the y-axis with the left hand then interchange them:



We face the opposite side of the paper then, but the graph is still visible: The x-axis is now pointing up and the y-axis right. The axes can also be *skewed*:



Even then, the two numbers indicating the intersection of two lines will unambiguously determine a location on the plane. And so on... Further analysis is presented in Volume 4 (Chapter 4DE-2).

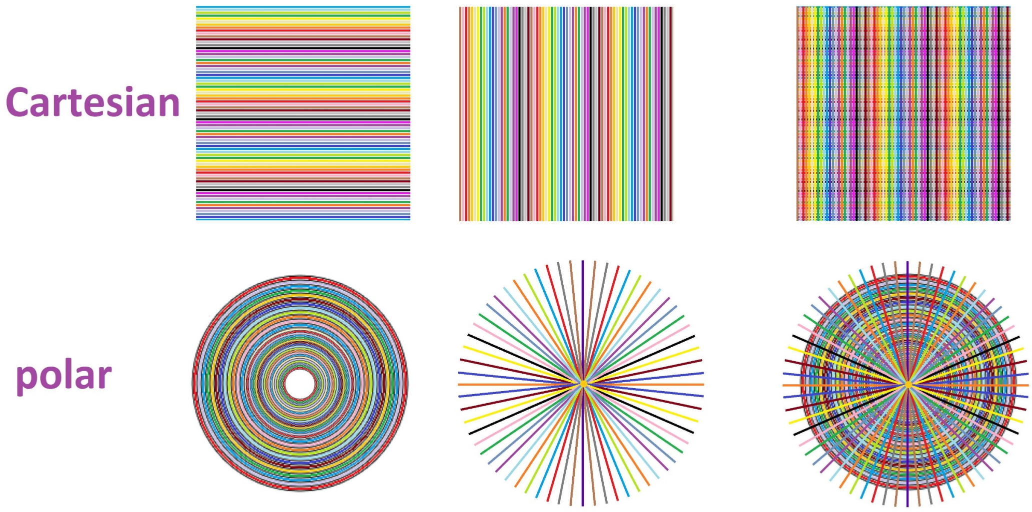
There are also *alternative coordinate systems*. Not only the axes are not rectangular; they are also *curved*!

The *circle* is a very special parametric curve. This curve will also supply us with a new way to record locations on the plane. It is called the *polar coordinate system*.

What makes coordinate systems possible is our ability to unambiguously assign each point to certain predetermined sets. Parallel lines don't intersect and they also cover the whole plane. That's why we have the following:

1. Every point belongs to one and only one horizontal line.
2. Every point belongs to one and only one vertical line.

Providing one from either class for each point is how the Cartesian system works:



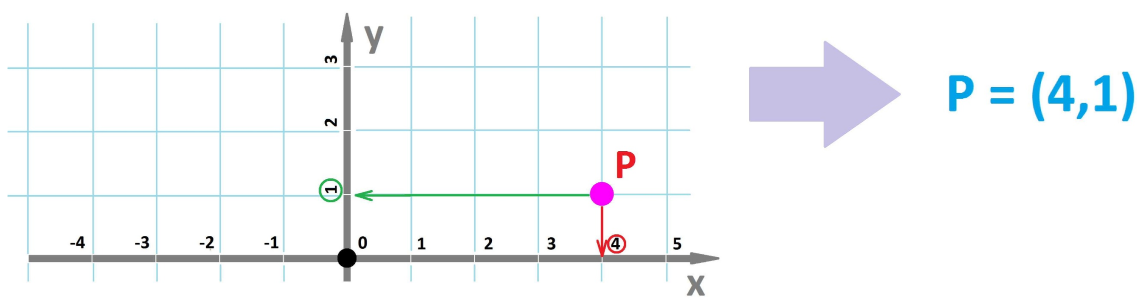
Now, we also have:

1. Concentric circles don't intersect and they, plus the center, also cover the whole plane. That's why every point belongs to one and only one of these circles, or the center.
2. Lines through the origin have only one point in common and they also cover the whole plane. That's why every point, other than the origin, belongs to one and only one of these lines.

Providing one from either class for each point is how the polar system works.

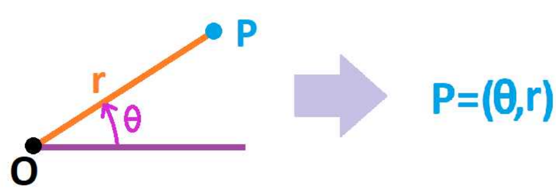
Now, the numerical representations of locations, i.e., their *coordinates*.

The Cartesian coordinate system is based on the idea of measuring the distances from the point to the axes.

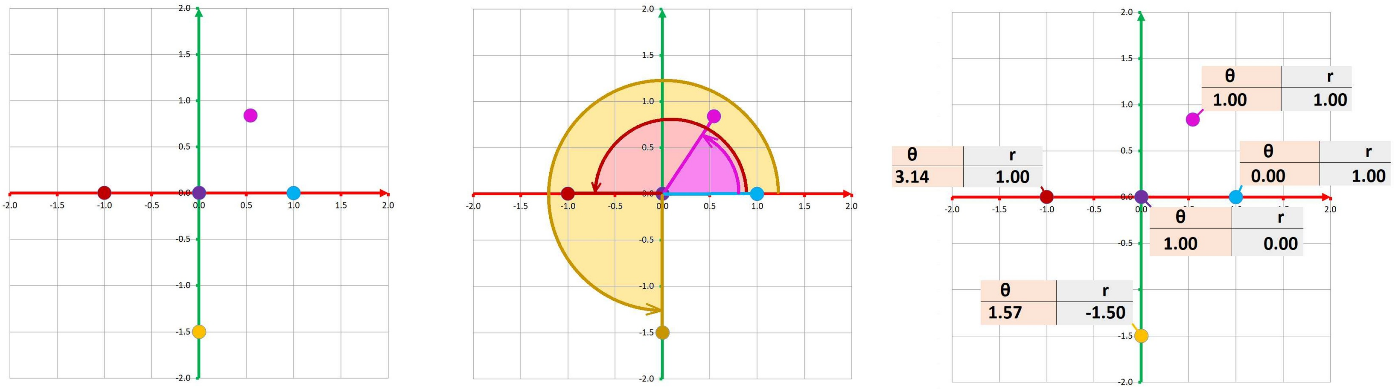


The polar coordinate system is based on these two ideas:

- measuring the distance from the point to the origin and
- measuring the angle with the x-axis.



For example, we have these:



The definition is independent from the presence of a Cartesian system:

Definition 4.6.2: polar coordinates

Suppose a point O called the *pole* and a ray L called the *polar axis* starting at O are given on the plane. Then for any point P, its *polar coordinates* are the two numbers, θ and r, defined as follows:

- θ is the angle from L to the line OP in the counterclockwise direction.
- r is the distance from P to O.

They commonly co-exist though.

Definition 4.6.3: associated polar and Cartesian systems

A polar and a Cartesian coordinate systems are called *associated* with each when the pole of the former is the origin of the latter and the polar axis of the former is the x-axis of the latter.

Warning!

No matter what θ is, it's O if $r = 0$.

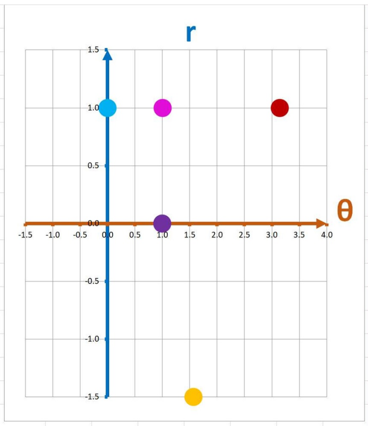
Example 4.6.4: polar points

Let's consider the relation between associated polar and Cartesian systems. For each of these pairs (θ, r), we compute its counterpart on the xy-plane:

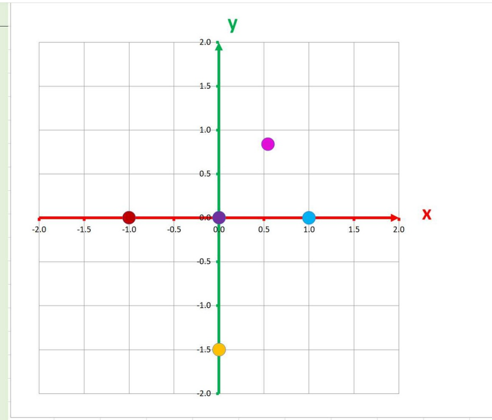
θ	r	$\to$	x	y
0.00	1.00	$\to$	1.00	0.00
1.00	0.00	$\to$	0.00	0.00
1.00	1.00	$\to$	0.54	0.84
3.14	1.00	$\to$	-1.00	0.00
1.57	-1.50	$\to$	0.00	-1.50

We plot them here:

θ	r
0.00	1.00
1.00	0.00
1.00	1.00
3.14	1.00
1.57	-1.50



x	y
1.00	0.00
0.00	0.00
0.54	0.84
-1.00	0.00
0.00	-1.50



We used the following formulas for the last two columns:

`=RC[-9]*COS(RC[-10])` and `=RC[-10]*SIN(RC[-11])`

Theorem 4.6.5: Conversion Between Polar And Cartesian

1. *A point P with the polar coordinates r and θ has the following coordinates in the associated Cartesian system:*

$$\boxed{x = r\cos\theta,\ y = r\sin\theta}$$

2. *A point P with the Cartesian coordinates $x \neq 0$ and y has the following coordinates in the associated polar system:*

$$\boxed{\theta = \arctan\left(\frac{y}{x}\right),\ r = \sqrt{x^2 + y^2}}$$

Exercise 4.6.6

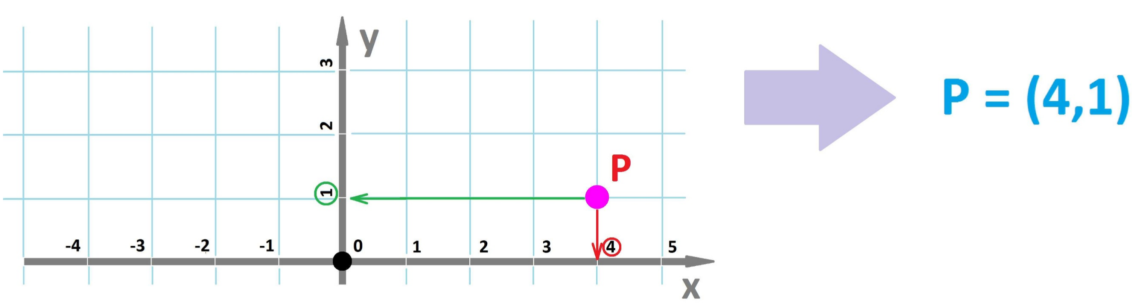
Prove the theorem.

The main difference from the Cartesian system is as follows.

The idea of a coordinate system is to associate a pair (or triple, etc.) of numbers to every location in an unambiguous way. It's a correspondence:

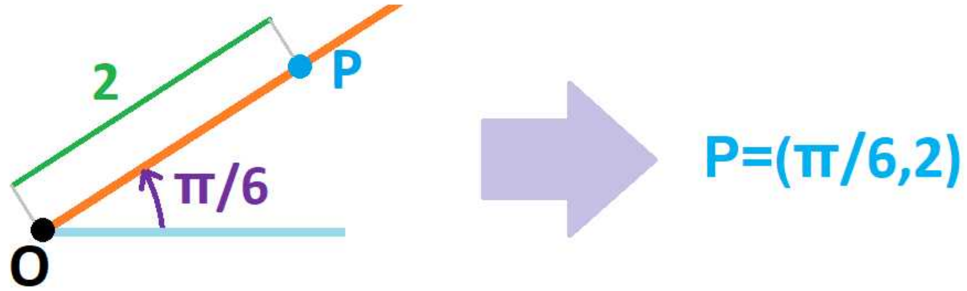
$$\boxed{\text{location} \quad \longleftrightarrow \quad \text{a pair of numbers}}$$

In the forward direction, $\boxed{\to}$. Cartesian:



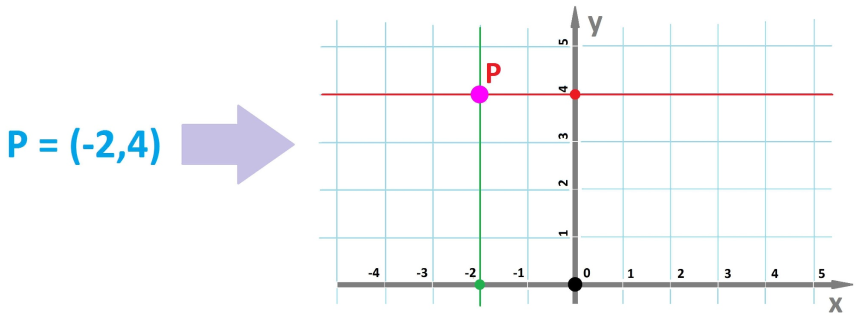
Suppose P is a location on the plane. We then draw a vertical line through P until it intersects the x-axis. The mark, x, of the location where they cross is the x-coordinate of P. We next draw a horizontal line through P until it intersects the y-axis. The mark, y, of the location where they cross is the y-coordinate of P. We end up with $r \geq 0$ and $0 \leq \theta < 2\pi$.

Polar:



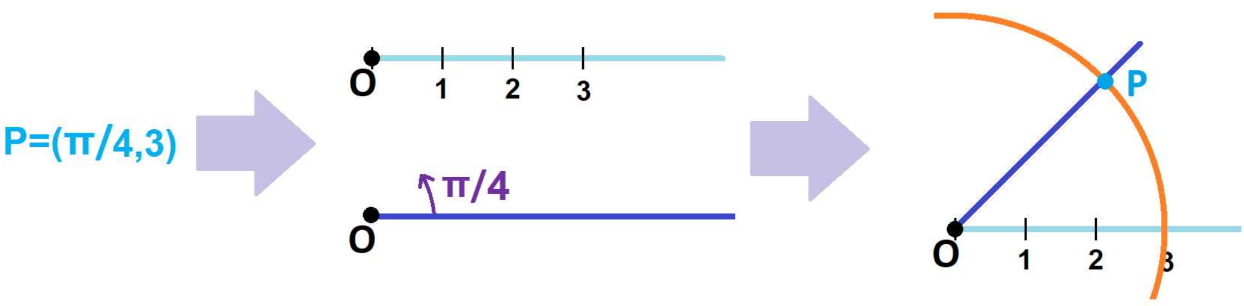
Suppose P is a location on the plane. We then draw a line from O through P. We measure the angle of OP with the polar axis counterclockwise. That's θ. We measure the distance from O to P. That's r.

In the backward direction, $\boxed{\leftarrow}$. Cartesian:



Suppose x and y are numbers. First, we find the mark x on the x-axis and draw a vertical line through this point. Second, we find the mark y on the y-axis and draw a horizontal line through this point. The intersection of these two lines is the corresponding location P on the plane.

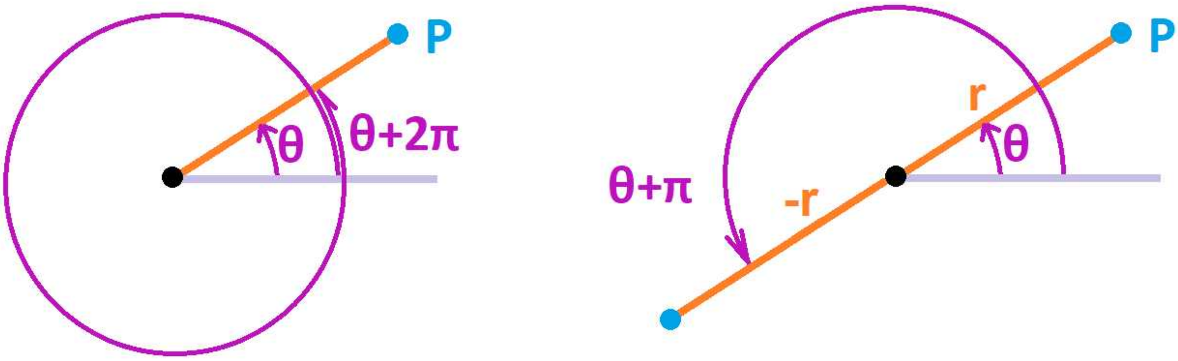
Polar:



Suppose θ and r are numbers. First, we rotate the polar axis θ radians in the counterclockwise direction. Second, we find the mark r on the polar axis and draw a circle centered at O through this point. The intersection of these two is the corresponding location P on the plane.

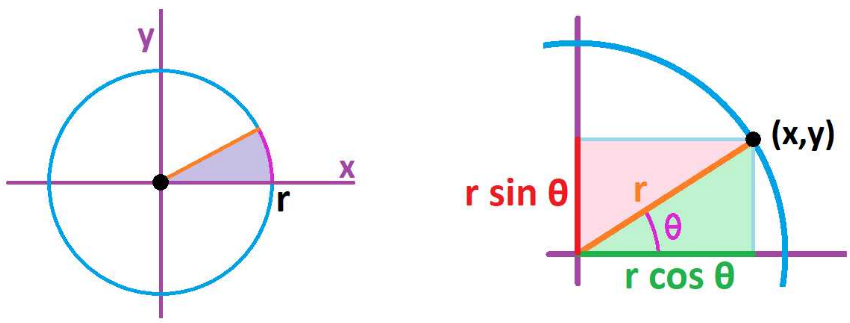
Unfortunately, there are more different pairs (θ, r) that produce (in addition to $(\theta, 0)$) identical locations. Unlike the Cartesian system, the polar coordinate system does not give an unambiguous representation of every location on the plane:

$$(\theta + 2\pi, r) = (\theta, r), \quad (\theta, -r) = (\theta + \pi, r).$$



We address this problem by using parametric curves.

Recall that the x-coordinate of the point on the circle of radius r at angle θ with the x-axis is $r\cos\theta$ and its y-coordinate is $r\sin\theta$:



Exercise 4.6.7

Suggest other examples of how two different polar coordinates produce the same point.

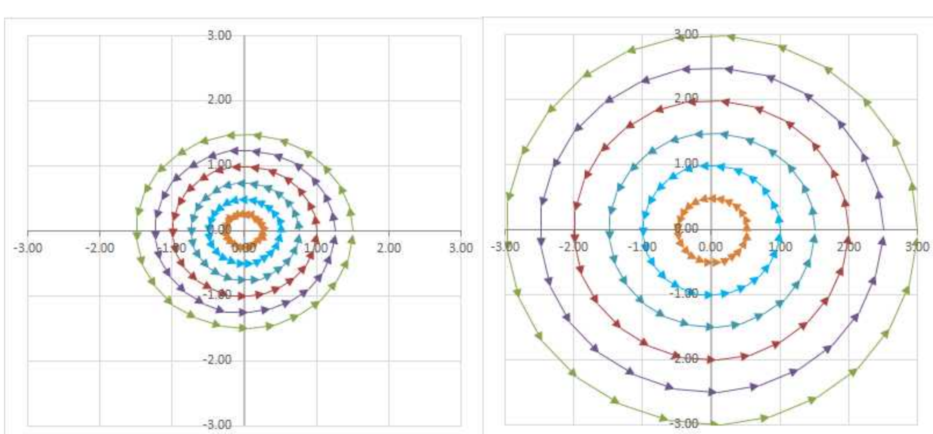
Exercise 4.6.8

Represent in polar coordinates these points given by their Cartesian coordinates: (a) $(1,2)$; (b) $(-1,-1)$; (c) $(0,0)$.

Exercise 4.6.9

Represent in Cartesian coordinates these points given by their polar coordinates: (a) $\theta=0,\ r=-1$; (b) $\theta=\pi/4,\ r=2$; (c) $\theta=1,\ r=0$.

We have a family of concentric circles parametrized in a uniform way:



Therefore, the pair of numbers,

$$\begin{array}{ccc} \theta & \text{and} & r \\ -\infty<\theta<+\infty & & -\infty<r<+\infty, \end{array}$$

determines (but not unambiguously) a location on the plane:

- The latter number, r, determines *which* circle we pick.
- The former number, θ, tells *how far* we go along this circle.

This ambiguity of the polar coordinate system allows us to effectively represent some complex curves.

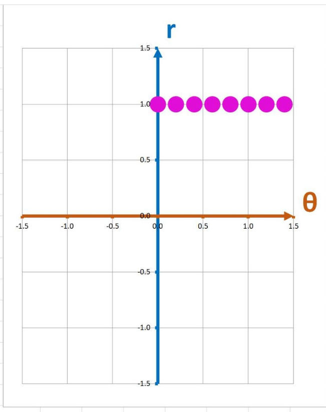
Example 4.6.10: polar curves

Simple relations between r and θ produce curves: on the θr-plane and on the xy-plane. The formula for the circle of radius R is very simple, by design:

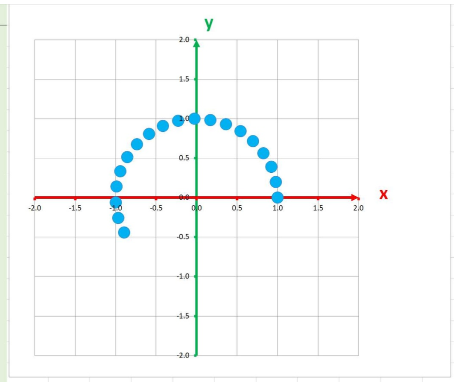
$$r = R\,.$$

So, r is fixed while θ varies:

θ	r
0.00	1.00
0.20	1.00
0.40	1.00
0.60	1.00
0.80	1.00
1.00	1.00
1.20	1.00
1.40	1.00
1.60	1.00
1.80	1.00
2.00	1.00
2.20	1.00
2.40	1.00
2.60	1.00
2.80	1.00
3.00	1.00
3.20	1.00
3.40	1.00
3.60	1.00

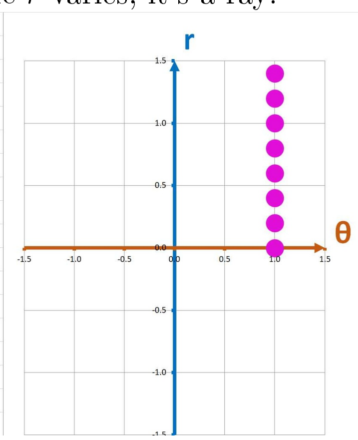


x	y
1.00	0.00
0.98	0.20
0.92	0.39
0.83	0.56
0.70	0.72
0.54	0.84
0.36	0.93
0.17	0.99
-0.03	1.00
-0.23	0.97
-0.42	0.91
-0.59	0.81
-0.74	0.68
-0.86	0.52
-0.94	0.33
-0.99	0.14
-1.00	-0.06
-0.97	-0.26
-0.90	-0.44

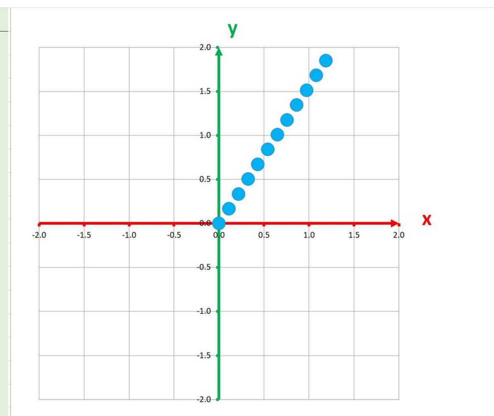


Next, θ is fixed while r varies; it's a ray:

θ	r
1.00	0.00
1.00	0.20
1.00	0.40
1.00	0.60
1.00	0.80
1.00	1.00
1.00	1.20
1.00	1.40
1.00	1.60
1.00	1.80
1.00	2.00
1.00	2.20
1.00	2.40
1.00	2.60
1.00	2.80
1.00	3.00
1.00	3.20

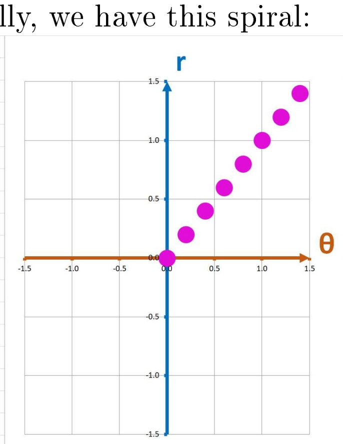


x	y
0.00	0.00
0.11	0.17
0.22	0.34
0.32	0.50
0.43	0.67
0.54	0.84
0.65	1.01
0.76	1.18
0.86	1.35
0.97	1.51
1.08	1.68
1.19	1.85
1.30	2.02
1.40	2.19
1.51	2.36
1.62	2.52
1.73	2.69

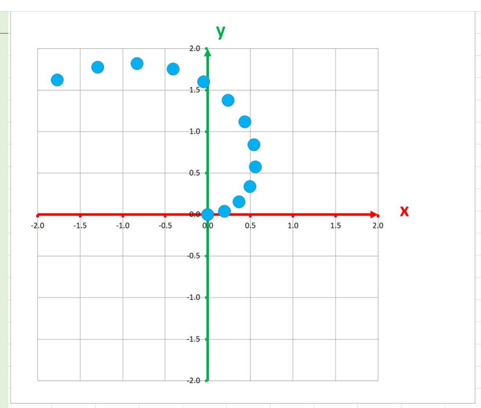


If both vary, identically, we have this spiral:

θ	r
0.00	0.00
0.20	0.20
0.40	0.40
0.60	0.60
0.80	0.80
1.00	1.00
1.20	1.20
1.40	1.40
1.60	1.60
1.80	1.80
2.00	2.00
2.20	2.20
2.40	2.40
2.60	2.60
2.80	2.80
3.00	3.00
3.20	3.20
3.40	3.40

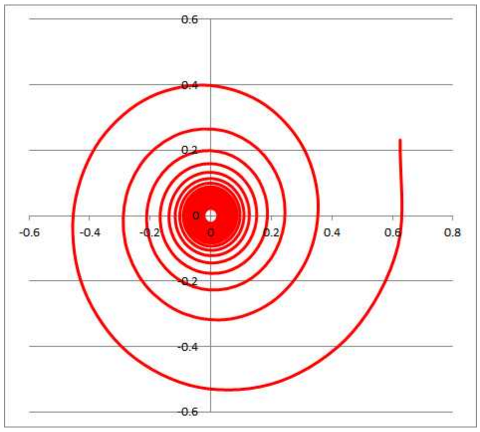


x	y
0.00	0.00
0.20	0.04
0.37	0.16
0.50	0.34
0.56	0.57
0.54	0.84
0.43	1.12
0.24	1.38
-0.05	1.60
-0.41	1.75
-0.83	1.82
-1.29	1.78
-1.77	1.62
-2.23	1.34
-2.64	0.94
-2.97	0.42
-3.19	-0.19
-3.29	-0.87



Example 4.6.11: spiral

Suppose we need to find a polar representation of this spiral, winding onto the origin:

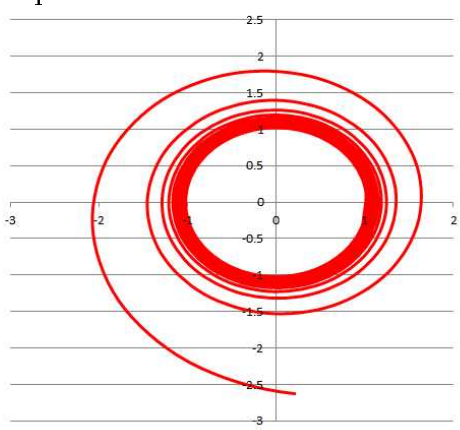


We realize that we need r to approach 0 as θ goes to infinity. For example, this will do:

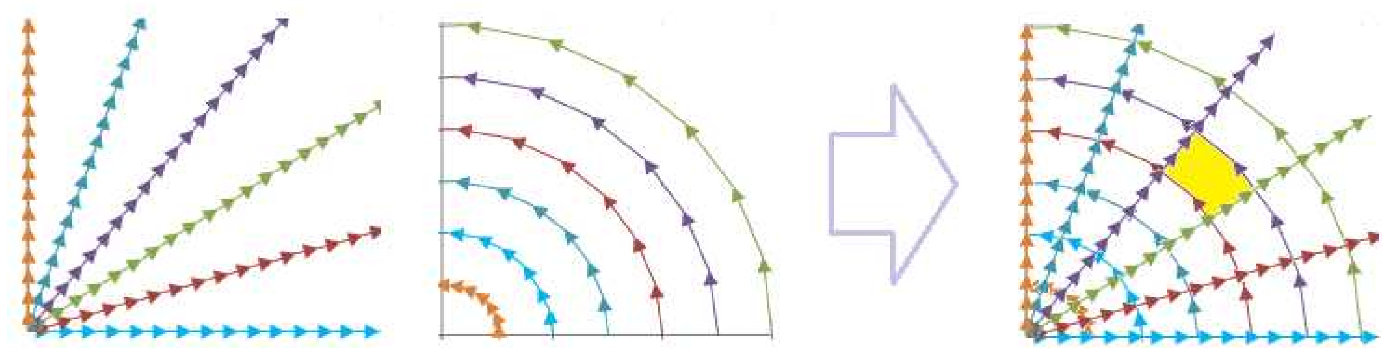
$$r = 1/\theta, \quad \theta > 0\,.$$

Exercise 4.6.12

Find a polar representation of this spiral:



The rectangular grid of the Cartesian system is replaced with this:



Exercise 4.6.13

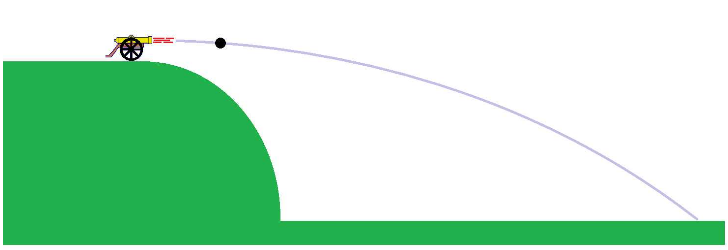
Represent one of the curved rectangles as a set using polar coordinates.

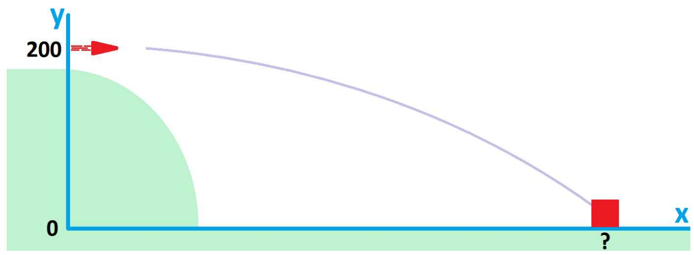
4.7. Vectors on the plane

Let's take another look at the problem from the beginning of the chapter.

Problem: From a 200-feet elevation, a cannon is fired horizontally at 200 feet per second. How far will the

cannonball go?





The result was these recursive formulas computed as the Riemann sums:

	horizontal	vertical
acceleration	a_n	b_n
velocity	$v_{n+1} = v_n + ha_n$	$u_{n+1} = u_n + hb_n$
position	$x_{n+1} = x_n + hv_n$	$y_{n+1} = y_n + hu_n$

The next step is obvious:

▶ We interpret the locations as *points* on the plane.

They are computed recursively, just as their coordinates are:

location		velocity		displacement	
horizontal	vertical	horizontal	vertical	horizontal	vertical
$(x_n\,,$	$y_n)$	$< v_n\,,$	$u_n >$	$< hv_n\,,$	$hu_n >$

The next question is then:

▶ How do we interpret the velocities and the displacements?

They are *vectors*.

Just as we combine pairs of numbers into points for locations,

$$P_n = (x_n, y_n)\,,$$

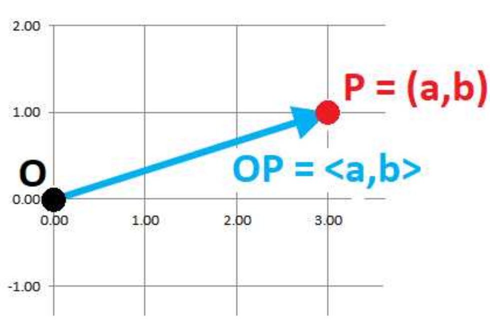
we now combine the pairs of numbers into vectors for velocities and displacements:

$$V_n =< v_n, u_n >,\ D_n =< hv_n, hu_n > \ .$$

The difference is in algebra... But first the definitions.

Definition 4.7.1: vector in dimension 2

If a segment's starting point is the origin, i.e., it's OP for some P, it is called a (2-dimensional) *vector* in $\mathbf{R}^2$.



Definition 4.7.2: vector in xy-plane

The *components* of vector OP are the coordinates of its terminal point P, according to the following notation:

$$\boxed{P = (a, b) \iff OP =< a, b >}$$

Warning!

It is also common to use (a, b) to denote the vector.

Let's consider the algebra. This is what we have when we treat the two directions separately:

$$x_{n+1} = x_n + hv_n,\ y_{n+1} = y_n + hu_n\,.$$

After combining these pairs of numbers into points and into vectors, we have:

$$(x_{n+1}, y_{n+1}) = (x_n, y_n) + < hv_n, hu_n > \ .$$

The addition of points and vectors is discussed in Volume 4. We concentrate on the following two operations.

First, the displacements are the velocities multiplied by the time increment h:

$$D_n = hV_n\,.$$

This is *scalar multiplication of vectors*.

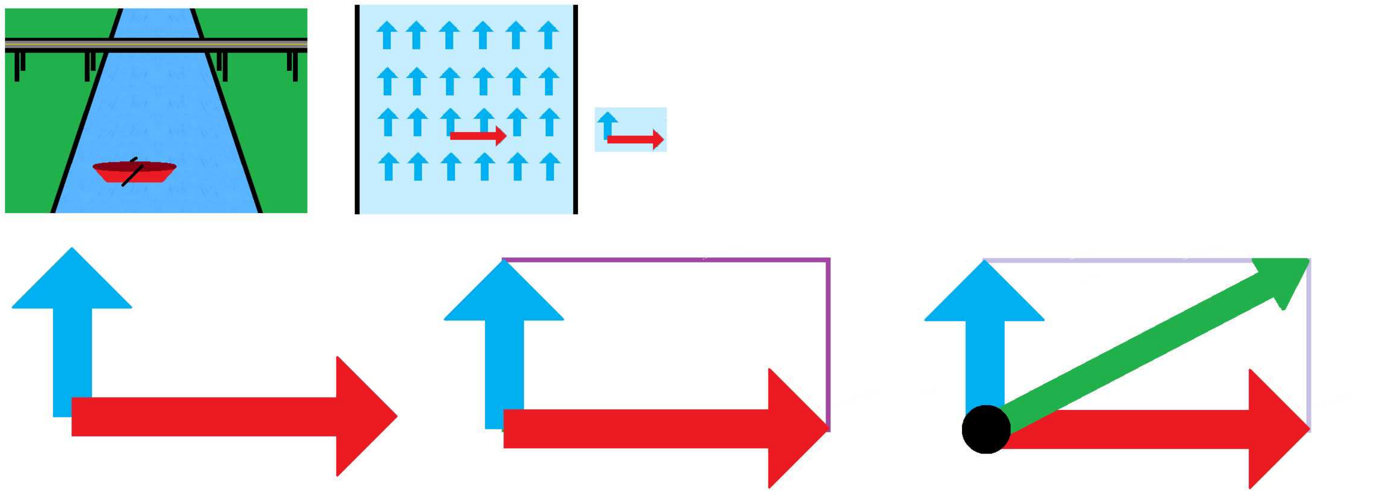
Second, the displacements are added consecutively:

$$P_{n+1} = P_n + D_n = (P_{n-1} + D_{n-1}) + D_n = P_{n-1} + (D_{n-1} + D_n)\,.$$

This is *vector addition*.

Example 4.7.3: velocity of stream

If we look at the velocities of particles in a stream, they may also be combined with the speed of rowing of the boat:

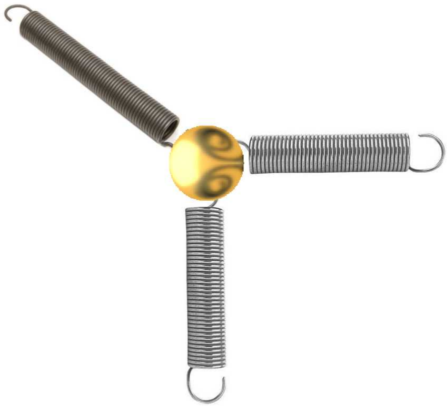


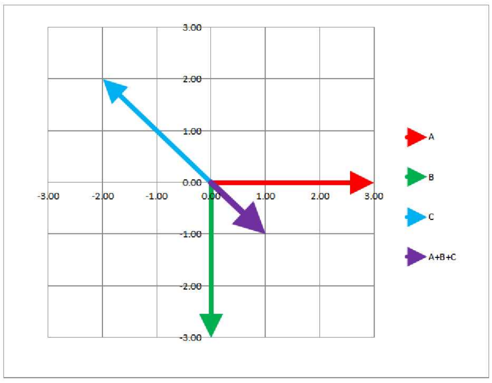
Exercise 4.7.4

With the velocities as shown, what is the best strategy to cross the canal?

Example 4.7.5: forces

Let's also look at *forces as vectors*. For example, springs attached to an object will pull it in their respective directions:





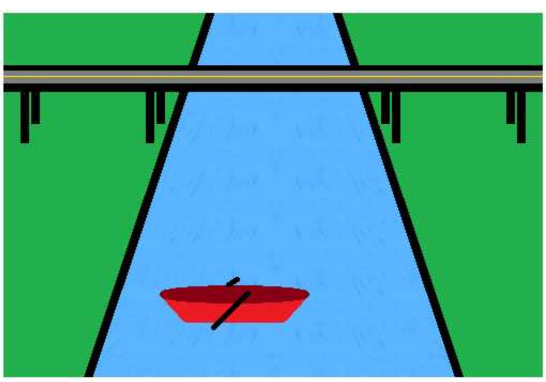
We add these vectors to find the combined force as if produced by a single spring. The forces are vectors that start at the same location.

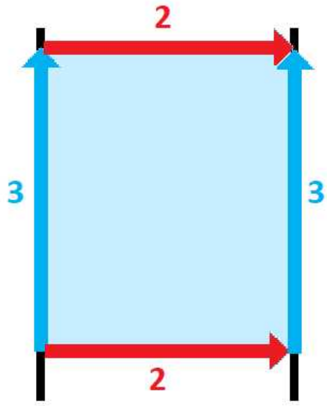
Example 4.7.6: displacements

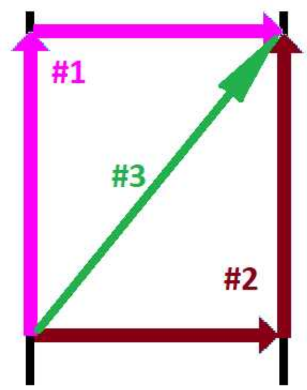
We can interpret the displacements, too, as vectors aligned to their starting points. Imagine we are crossing a river 3 miles wide and we know that the current takes us 2 miles downstream. There are three different ways this can happen:

1. a trip 3 miles north followed by a trip 2 miles east; or
2. a trip 2 miles east followed by a trip 3 miles north; but also
3. a trip along the diagonal of a rectangle with one side going 3 miles north and another 2 miles east.

The three outcomes are the same:







They are the same.

This is *vector algebra* we need to learn.

So, we have a coordinatewise addition:

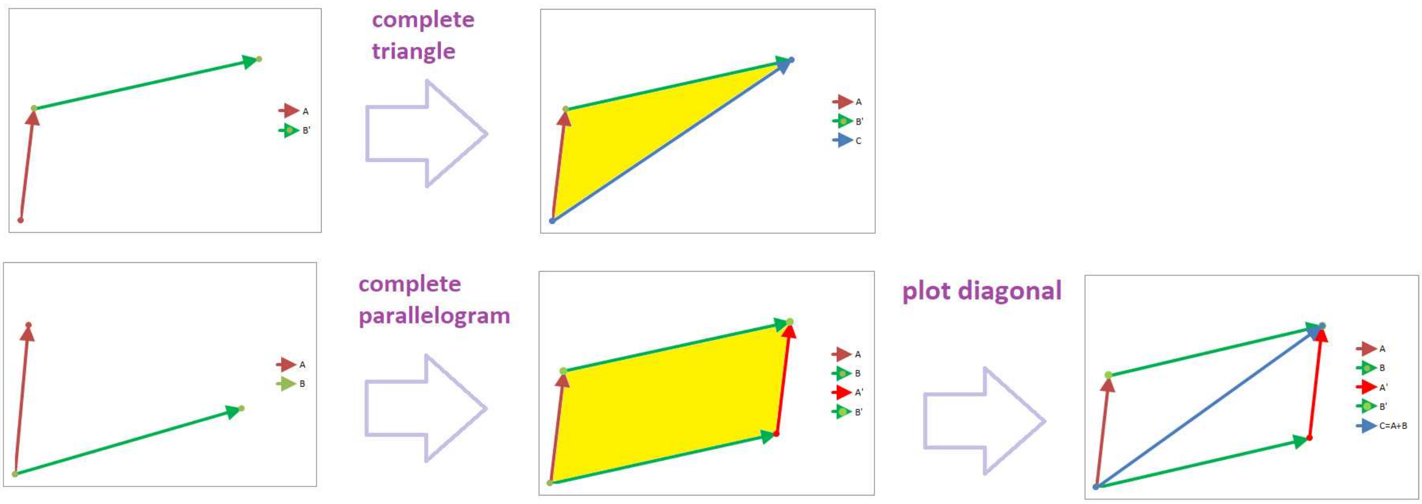
Here:

$$< 1, 2 > + < 2, -1 > = < 1 + 2, 2 + (-1) > = < 3, 1 > \ .$$

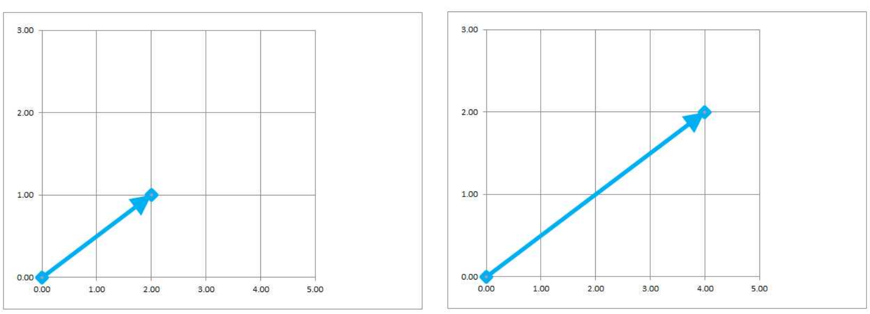
Geometrically, to add two vectors, we follow either:

1. The head-to-tail: the triangle construction.
2. The tail-with-tail: the parallelogram construction.

They have to produce the same result! They do, as illustrated below:



We also have a multiplication of the coordinates by the same number:



Here:

$$2 \cdot < 2, 1 > = < 2 \cdot 2, 2 \cdot 1 > = < 4, 2 > \ .$$

In summary, we have the following:

Definition 4.7.7: vector operations

1. Any two vectors $< a, b >$ and $< u, v >$ can be added, producing a new vector called their *sum*:

$$\boxed{< a, b > + < u, v > = < a + u, b + v >}$$

2. Any vector $< a, b >$ can be multiplied by a number k, producing a new vector called the *scalar product*:

$$\boxed{k < a, b > = < ka, kb >}$$

Along this algebra, there is sill *geometry* too.

A vector has a direction, which is one of the two directions of the line it determines, and a magnitude, defined as follows.

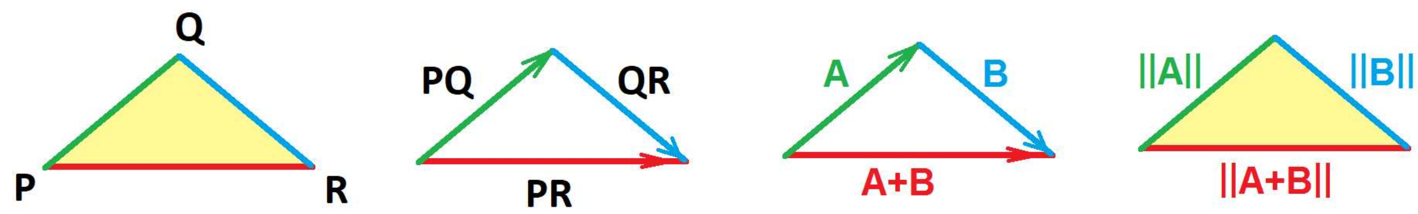
Definition 4.7.8: magnitude of vector

The *magnitude* of a vector $OP =< a, b >$ is defined as the distance from O to its tip P, denoted as follows:

$$\boxed{|| < a, b > || = \sqrt{a^2 + b^2}}$$

Algebra and geometry of vectors can interact. For example, the *Triangle Inequality* take this form:

$$\boxed{||A + B|| \leq ||A|| + ||B||}$$



Exercise 4.7.9

Why not "$<$"?

Exercise 4.7.10

Expand: $||kA|| = ...$

4.8. *How complex numbers emerge*

The equation

$$x^2 + 1 = 0$$

has no solutions. Indeed, we observe the following:

$$x^2 \geq 0 \implies x^2 + 1 > 0 \implies x^2 + 1 \neq 0\,.$$

If we try to solve it the usual way, we get these:

$$x = \sqrt{-1} \ \text{ and } \ x = -\sqrt{-1}\,.$$

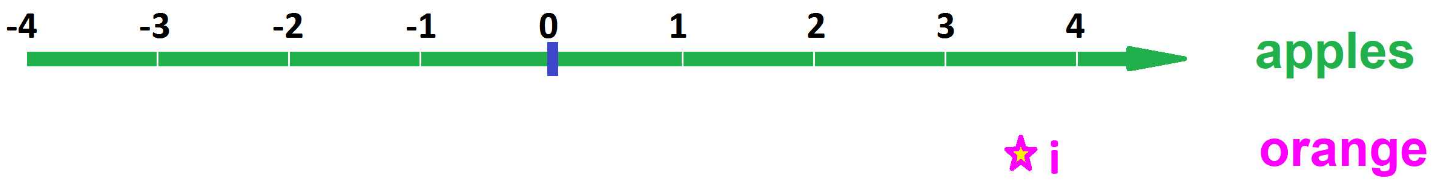
There are no such *real numbers*.

However, let's ignore this fact for a moment. Let's substitute what we have back into the equation and – blindly – follow the rules of algebra. We "confirm" that this "number" is a "solution":

$$x^2 + 1 = (\sqrt{-1})^2 + 1 = (-1) + 1 = 0\,.$$

We call this entity the *imaginary unit*, denoted by i.

We just add this "number" to the set of numbers we do algebra with:



And see what happens...

Making i a part of algebra will only require this three-part convention:

1. i is not a real number (and, in particular, $i \neq 0$), but
2. i can participate in the (four) algebraic operations with real numbers by following the same rules; also
3. $i^2 = -1$.

What algebraic rules are those? A few very basic ones:

$$x + y = y + x,\ x \cdot y = y \cdot x,\ x(y + z) = xy + xz,\ \text{etc.}$$

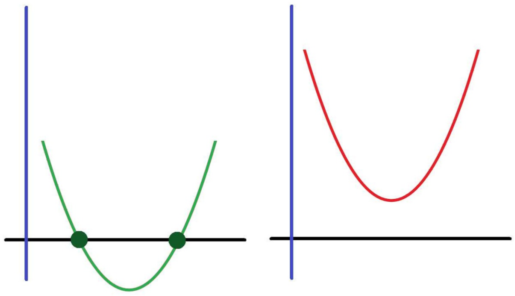
We allow one or several of these parameters to be i. For example, we have:

$$i + y = y + i,\ i \cdot y = y \cdot i,\ i(y + z) = iy + iz,\ \text{etc.}$$

What makes this extra effort worthwhile is a new look at quadratic polynomials. For example, this is how we may factor one:

$$x^2 - 1 = (x - 1)(x + 1)\,.$$

Then $x = 1$ and $x = -1$ are the x-intercepts of the polynomial:



But some polynomials, called *irreducible*, cannot be factored; there are no a, b such that:

$$x^2 + 1 = (x - a)(x - b)\,.$$

There are no *real* a, b, that is! Using our rules, we discover:

$$(x - i)(x + i) = x^2 - ix + ix - i^2 = x^2 + 1\,.$$

Of course, the number i is *not* an x-intercept of $f(x) = x^2 + 1$ as the x-axis ("the real line") consists of only (and all) real numbers.

So, multiples of i appear immediately as we start doing algebra with it.

Definition 4.8.1: imaginary numbers

The real multiples of the imaginary unit, i.e.,

$$z = ri,\ r \text{ real},$$

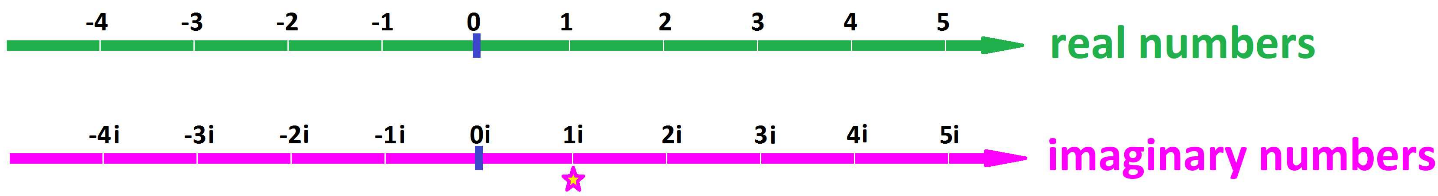
are called *imaginary numbers*.

We have created a whole class of *non-real* numbers! Of course, ri, where r is real, can't be real:

$$(ri)^2 = r^2 i^2 = -r^2 < 0\,.$$

The only exception is $0i = 0$; it's real!

There are as many of them as the real numbers:



Example 4.8.2: quadratic equations

The imaginary numbers may also come from solving the simplest quadratic equations. For example, the equation

$$x^2 + 4 = 0$$

gives us via our substitution:

$$x = \pm\sqrt{-4} = \pm\sqrt{4(-1)} = \pm\sqrt{4}\sqrt{-1} = \pm 2i\,.$$

Indeed, if we substitute $x = 2i$ into the equation, we have:

$$(2i)^2 + 4 = (2)^2(i)^2 + 4 = 4(-1) + 4 = 0\,.$$

More general quadratic equations are discussed in the next section.

Imaginary numbers obey the laws of algebra as we know them! If we need to simplify the expression, we try to manipulate it in such a way that real numbers are combined with real while i is pushed aside.

For example, we can just factor i out of all addition and subtraction:

$$5i + 3i = (5+3)i = 8i\,.$$

It looks exactly like middle school algebra:

$$5x + 3x = (5+3)x = 8x\,.$$

After all, x *could* be i. Another similarity is with the algebra of quantities that have units:

$$5 \text{ in.} + 3 \text{ in.} = (5+3) \text{ in.} = 8 \text{ in.}\,.$$

So, the nature of the unit doesn't matter (if we can push it aside). Even simpler:

$$5 \text{ apples } + 3 \text{ apples } = (5+3) \text{ apples } = 8 \text{ apples }\,.$$

It's "8 apples" not "8"! And so on.

This is how we multiply an imaginary number by a real number:

$$2 \cdot (3i) = (2 \cdot 3)i = 6i\,.$$

We have a new imaginary number.

How do we multiply two imaginary numbers? It's different; after all, we don't usually multiply apples by apples! In contrast to the above, even though multiplication and division follow the same rule as always, we can, when necessary, and often have to, simplify the outcome of our algebra using our *fundamental identity*:

$$i^2 = -1\,.$$

For example:

$$(5i) \cdot (3i) = (5 \cdot 3)(i \cdot i) = 15i^2 = 15(-1) = -15\,.$$

It's real!

We also simplify the outcome by using the other *fundamental fact* about the imaginary unit:

$$i \neq 0\,.$$

We can divide by i! For example,

$$\frac{5i}{3i} = \frac{5}{3}\frac{i}{i} = \frac{5}{3} \cdot 1 = \frac{5}{3}\,.$$

As you can see, doing algebra with imaginary numbers will often bring us back to real numbers. These two classes of numbers cannot be separated from each other!

They aren't. Let's take another look at quadratic equations. The equation

$$ax^2 + bx + c = 0,\ a \neq 0\,,$$

is solved with the familiar *Quadratic Formula*:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}\,.$$

Let's consider

$$x^2 + 2x + 10 = 0\,.$$

Then the roots are supposed to be:

$$\begin{aligned} x &= \frac{-2 \pm \sqrt{2^2 - 4 \cdot 10}}{2} \\ &= \frac{-2 \pm \sqrt{-36}}{2} \\ &= -1 \pm \sqrt{-9} && \text{There is no real solution!} \\ &= -1 \pm \sqrt{9}\sqrt{-1} && \text{But we go on.} \\ &= -1 \pm 3i\,. \end{aligned}$$

We end up adding real and imaginary numbers!

As there is no way to simplify this, we conclude the following:

- ▶ A number $a + bi$, where $a, b \neq 0$ are real, is neither real nor imaginary.

Exercise 4.8.3

Explain why.

This addition is not literal. It's like "adding" apples to oranges:

$$5 \text{ apples } + 3 \text{ oranges } = \ldots$$

It's not 8 and it's not 8 fruit because we wouldn't be able to read this equality backwards. The algebra will, however, be meaningful:

$$(5a + 3o) + (2a + 4o) = (5 + 3)a + (3 + 4)o = 8a + 7o\,.$$

It is as if we collect *similar terms*, like this:

$$(5 + 3x) + (2 + 4x) = (5 + 2) + (3 + 4)x = 8 + 7x\,.$$

This idea enables us to do this:

$$(5+3i)+(2+4i)=(5+3)+(3+4)i=8+7i\,.$$

Each of the numbers we are facing contain both real numbers and imaginary parts. This fact makes them "complex"...

Definition 4.8.4: complex number

Any sum of real and imaginary numbers is called a *complex number*. The set of all complex numbers is denoted as follows:

$$\boxed{\mathbf{C}=\{z=a+bi:\ a,b \text{ real}\}}$$

Warning!

All real numbers are complex ($b=0$).

Addition and subtraction are easy; we just combine *similar terms* just like in middle school. For example,

$$(1+5i)+(3-i)=1+5i+3-i=(1+3)+(5i-i)=4+4i\,.$$

To simplify *multiplication* of complex numbers, we expand and then use $i^2=-1$, as follows:

$$\begin{aligned}(1+5i)\cdot(3-i)\ &=1\cdot 3+5i\cdot 3+1\cdot(-i)+5i\cdot(-i)\\ &=3+15i-i-5i^2\\ &=(3+5)+(15i-i)\\ &=8+14i\,.\end{aligned}$$

It's a bit trickier with *division*:

$$\begin{aligned}\frac{1+5i}{3-i}\ &=\frac{1+5i}{3-i}\frac{3+i}{3+i}\\ &=\frac{(1+5i)(3+i)}{(3-i)(3+i)}\\ &=\frac{-2+8i}{3^2-i^2}\\ &=\frac{-2+8i}{3^2+1}\\ &=\frac{1}{10}(-2+8i)\\ &=-0.2+0.8i\,.\end{aligned}$$

The simplification of the denominator is made possible by the trick of multiplying by $3+i$. It is the same trick we used in Volume 1 to simplify fractions with roots to compute their limits:

$$\frac{1}{1-\sqrt{x}}=\frac{1}{1-\sqrt{x}}\frac{1+\sqrt{x}}{1+\sqrt{x}}=\frac{1+\sqrt{x}}{1-x}\,.$$

Definition 4.8.5: complex conjugate

The *complex conjugate* of $z=a+bi$ is defined and denoted as follows:

$$\bar{z}=\overline{a+bi}=a-bi\,.$$

The following is crucial.

Theorem 4.8.6: Algebra of Complex Numbers

The rules of the algebra of complex numbers are identical to those of real numbers:

- ***Commutativity of addition:*** $z + u = u + z$
- ***Associativity of addition:*** $(z + u) + v = z + (u + v)$
- ***Commutativity of multiplication:*** $z \cdot u = u \cdot z$
- ***Associativity of multiplication:*** $(z \cdot u) \cdot v = z \cdot (u \cdot v)$
- ***Distributivity:*** $z \cdot (u + v) = z \cdot u + z \cdot v$

This is the *complex number system*; it follows the rules of the real number system but also contains it. This theorem will allow us to build calculus for complex functions that is almost identical to calculus for real functions but also contains it.

Definition 4.8.7: standard form of complex number

Every complex number x has the *standard representation*:

$$z = a + bi\,,$$

where a and b are two real numbers. The two components are named as follows:

- a is the *real part* of z, with notation:

$$a = \mathrm{Re}(z)\,;$$

- bi is the *imaginary part* of z, with notation:

$$b = \mathrm{Im}(z)\,.$$

Then, the purpose of the computations above was to find the standard form of a complex number that comes from algebraic operations with other complex numbers. They were literally simplifications.

The definition makes sense because of the following result:

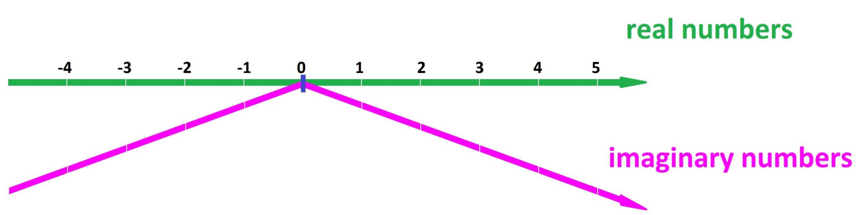
Theorem 4.8.8: Standard Form of Complex Number

Two complex numbers are equal if and only if both their real and imaginary parts are equal.

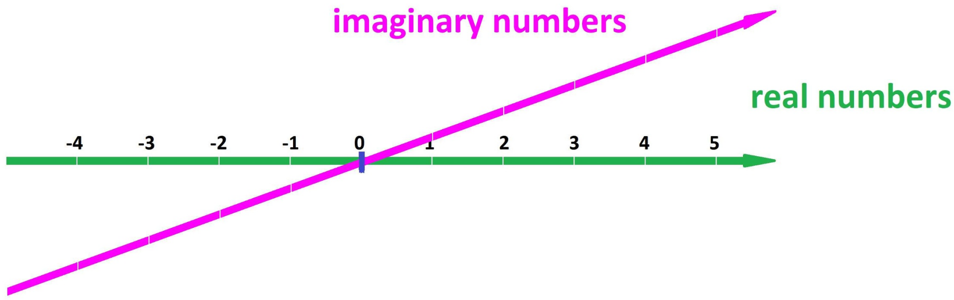
So, we have:

$$z = \mathrm{Re}(z) + \mathrm{Im}(z)i\,.$$

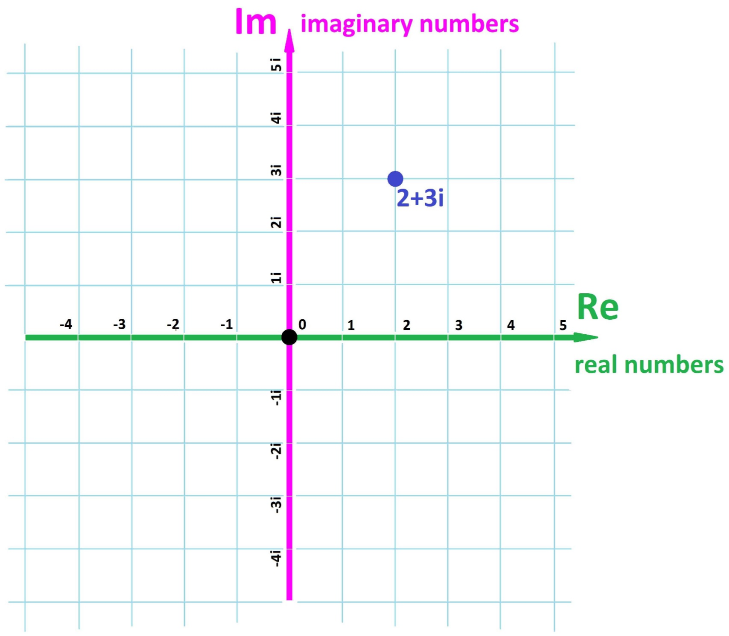
In order to see the *geometric representation of complex numbers*, we need to combine the real number line and the imaginary number line. How? We realize that they have nothing in common... except $0 = 0i$ belongs to both:



We can try to combine them like that, or like this:



Or we can try to combine them in the same manner we built the xy-plane:



This representation helps us understand the main idea:

- ▶ Complex numbers are linear combinations of the real unit, 1, and the imaginary unit, i.

If $z = a + bi$, then a and b are thought of as the components of vector z in the plane. We have a one-to-one correspondence:

$$\mathbf{C} \longleftrightarrow \mathbf{R}^2 \,,$$

given by

$$a + bi \ \longleftrightarrow\ < a, b > \ .$$

Then the x-axis of this plane consists of the real numbers and the y-axis of the imaginary numbers.

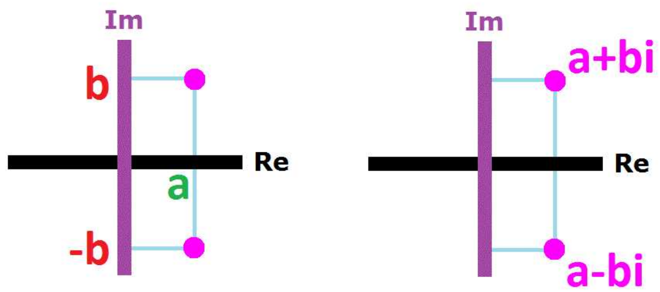
It is called the *complex plane*.

Warning!

This is just a visualization.

Then the complex conjugate of z is the complex number with the same real part as z and the imaginary part with the opposite sign:

$$\mathrm{Re}(\bar{z}) = \mathrm{Re}(z) \quad \text{and} \quad \mathrm{Im}(\bar{z}) = -\,\mathrm{Im}(z)\,.$$



Warning!

All numbers we have encountered so far are real non-complex, and so are all quantities one can encounter in day-to-day life or science: time, location, length, area, volume, mass, temperature, money, etc.

4.9. The complex plane C is the Euclidean space $\mathbf{R}^2$

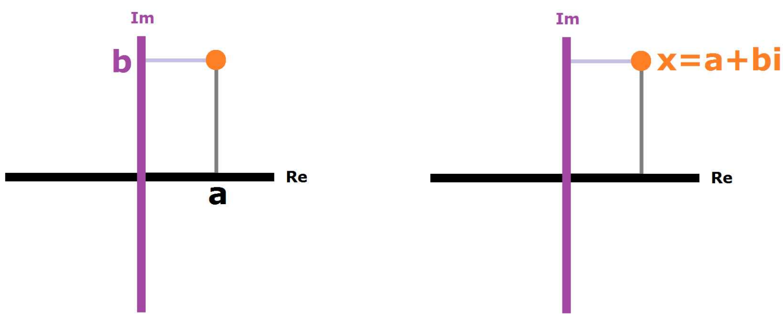
If we call complex number *numbers*, they must be subject to some *algebraic operations*.

We will initially look at them through the lens of *vector algebra* of the plane $\mathbf{R}^2$.

A complex number z has the *standard representation*:

$$z = a + bi\,,$$

where a and b are two real numbers. These two can be seen in the *geometric representation* of complex numbers:



Therefore, a and b are thought of as the coordinates of z as a *point* on the plane. But any complex number is not only a point on the complex plane but also a *vector*. We have a correspondence:

$$\mathbf{C} \longleftrightarrow \mathbf{R}^2\,,$$

given by

$$\boxed{a + bi \ \longleftrightarrow\ < a, b >}$$

There is more to this than just a match; the algebra of vectors in $\mathbf{R}^2$ applies!

Warning!

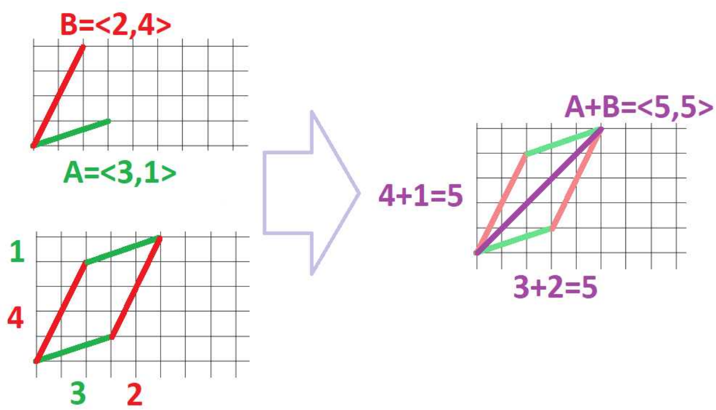
In spite of this fundamental correspondence, we will continue to think of complex numbers as *numbers* (and use the lower case letters).

Let's see how this algebra of numbers works in parallel with the algebra of 2-vectors.

First, the addition of complex numbers is done *componentwise*:

$$\begin{array}{ccccccccc}(a+bi) & + & (c+di) & = & (a+c) & + & (b+d)i \\ < a,b > & + & < c,d > & = & < a+c & , & b+d >\end{array}$$

It corresponds to addition of vectors:



Second, we can easily multiply complex numbers by real ones:

$$\begin{array}{cccccc}(a+bi) & c & = & (ac) & + & (bc)i \\ < a,b > & c & = & < ac & , & bc >\end{array}$$

It corresponds to scalar multiplication of vectors.

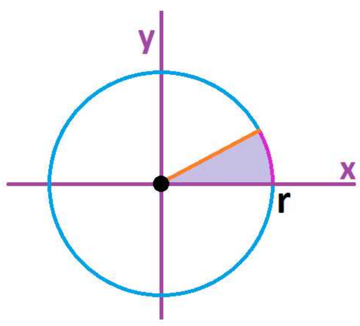
Warning!

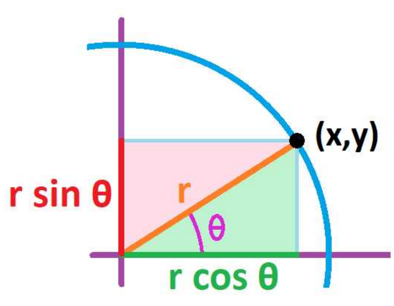
Vector algebra of $\mathbf{R}^2$ is complex algebra, but not vice versa. Complex *multiplication* is what makes it different.

Example 4.9.1: circle

We can easily represent circles on the complex plane:

$$z = r\cos\theta + r\sin\theta \cdot i\,.$$





Our study of calculus of complex numbers starts with the study of the *topology* of the complex plane. This topology is the same as that of *the Euclidean plane* $\mathbf{R}^2$!

Just as before, every function $z = f(t)$ with an appropriate domain creates a sequence:

$$z_k = f(k)\,.$$

A function with complex values defined on a ray in the set of integers, $\{p, p+1, ...\}$, is called an *infinite sequence*, or simply sequence.

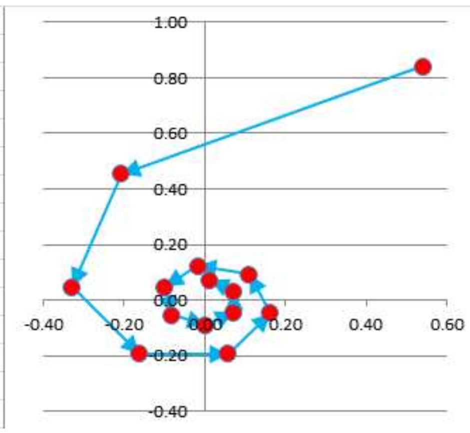
Example 4.9.2: spiral

A good example is that of the sequence made of the reciprocals:

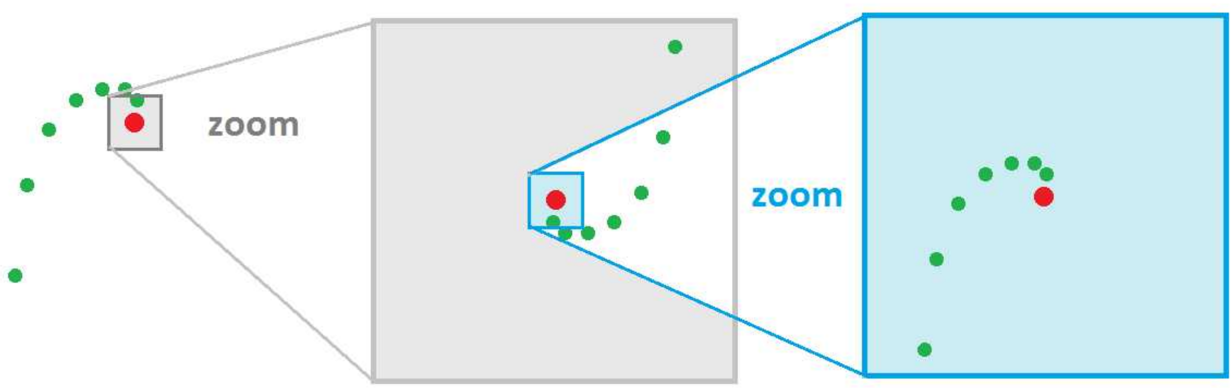
$$z_k = \frac{\cos k}{k} + \frac{\sin k}{k} i\,.$$

It *tends to* 0 while spiraling around it.

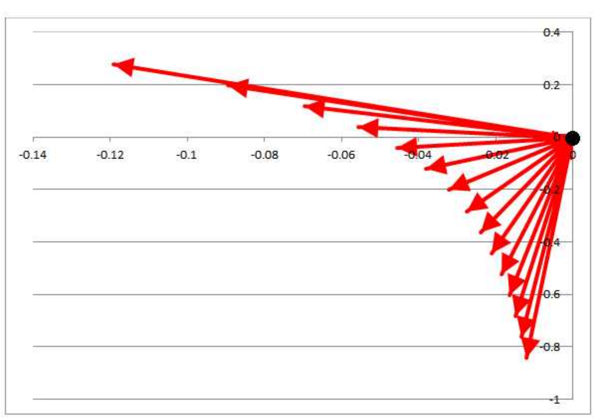
n	x	y
1.00	0.54	0.84
2.00	-0.21	0.45
3.00	-0.33	0.05
4.00	-0.16	-0.19
5.00	0.06	-0.19
6.00	0.16	-0.05
7.00	0.11	0.09
8.00	-0.02	0.12
9.00	-0.10	0.05
10.00	-0.08	-0.05
11.00	0.00	-0.09
12.00	0.07	-0.04
13.00	0.07	0.03
14.00	0.01	0.07



The starting point of calculus of complex numbers is the following. The convergence of a *sequence of complex numbers* is the convergence of its real and imaginary parts or, which is equivalent, the convergence of points (or vectors) on the complex plane seen as any plane: The distance from the kth point to the limit is getting smaller and smaller.



We use the definition of convergence for vectors on the plane by simply replacing vectors with complex numbers and "magnitude" with "modulus".



Definition 4.9.3: convergent sequence

Suppose $\{z_k : \ k = 1, 2, 3, \ldots\}$ is a sequence of complex numbers, i.e., points in $\mathbf{C}$. We say that the sequence *converges* to another complex number z, i.e., a point in $\mathbf{C}$, called the *limit* of the sequence, if:

$$||z_k - z|| \to 0 \text{ as } k \to \infty\,,$$

denoted as follows:

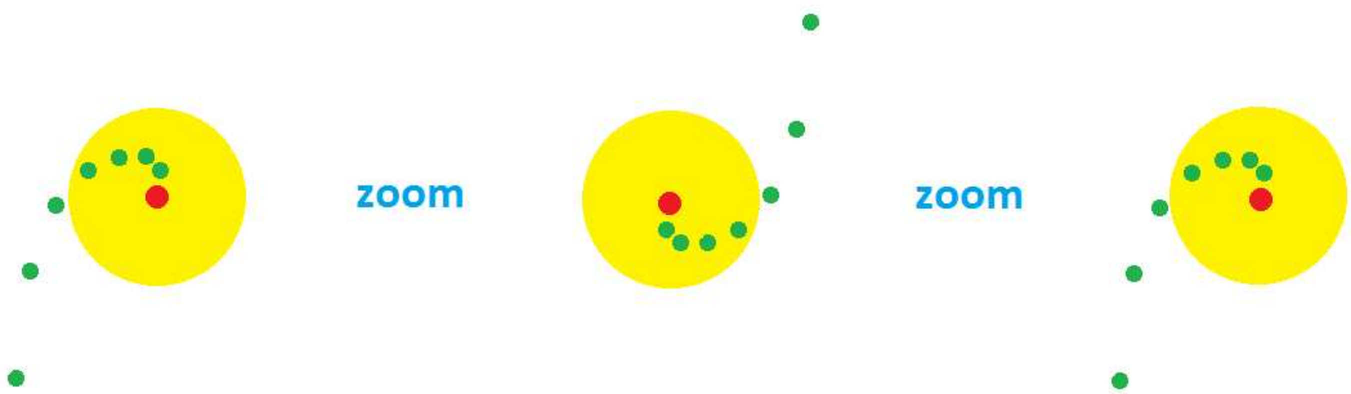
$$z_k \to z \text{ as } k \to \infty\,,$$

or

$$z = \lim_{k\to\infty} z_k\,.$$

If a sequence has a limit, we call the sequence *convergent* and say that it *converges*; otherwise it is *divergent* and we say it *diverges*.

In other words, the points start to accumulate in smaller and smaller circles around z. A way to visualize a trend in a convergent sequence is to enclose the tail of the sequence in a *disk*:



Theorem 4.9.4: Uniqueness of Limit

A sequence can have only one limit (finite or infinite); i.e., if a and b are limits of the same sequence, then $a = b$.

Definition 4.9.5: sequence tends to infinity

We say that a sequence z_k *tends to infinity* if the following condition holds: For each real number R, there exists such a natural number N that, for every natural number $k > N$, we have

$$||z_k|| > R\,.$$

We use the following notation:

$$z_k \to \infty \text{ as } k \to \infty\,.$$

The following is another analog of a familiar theorem about the topology of the plane.

Theorem 4.9.6: Componentwise Convergence of Sequences

A sequence of complex numbers z_k in $\mathbf{C}$ converges to a complex number z if and only if both the real and the imaginary parts of z_k converge to the real and the imaginary parts of z respectively; i.e.,

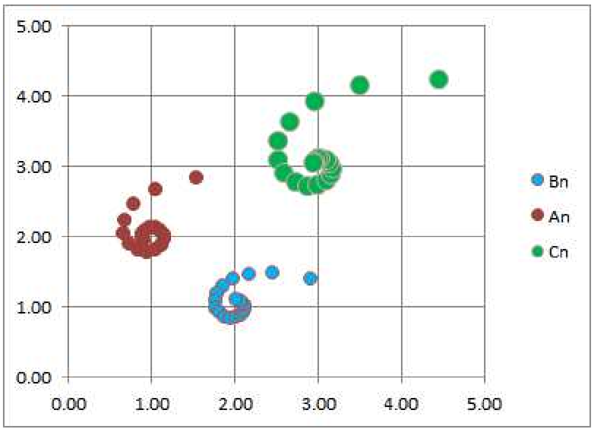
$$z_k \to z \iff \operatorname{Re}(z_k) \to \operatorname{Re}(z) \ \textit{ and } \ \operatorname{Im}(z_k) \to \operatorname{Im}(z)\,.$$

The algebraic properties of limits of sequences of complex numbers will also look familiar:

Theorem 4.9.7: Sum Rule for Complex Sequences

If sequences z_k, u_k converge, then so does $z_k + u_k$, and we have:

$$\lim_{k\to\infty}(z_k+u_k)=\lim_{k\to\infty} z_k+\lim_{k\to\infty} u_k\,.$$



Theorem 4.9.8: Constant Multiple Rule for Complex Sequences

If sequence z_k converges, then so does cz_k for any complex number c, and we have:

$$\lim_{k\to\infty} c\,z_k=c\cdot\lim_{k\to\infty} z_k\,.$$

Wouldn't calculus of complex numbers be just a copy of calculus on the plane? No, not with the possibility of *multiplication* taken into account.

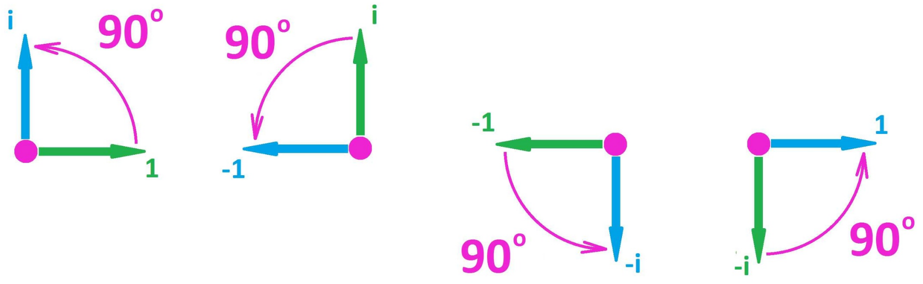
4.10. *Multiplication of complex numbers:* C *isn't just* R^2

So, the vector algebra of $\mathbf{R}^2$ is included in the complex algebra of $\mathbf{C}$. There is more to the latter.

Just like in $\mathbf{R}^2$, multiplication by a *real* number r will stretch/shrink all vectors and, therefore, the complex plane $\mathbf{C}$. However, multiplication by a *complex* number c will also rotate each vector.

Example 4.10.1: multiplication by i

Let's start with 1 and multiply it by i several times. Multiplication by i rotates the number by 90 degrees: 1 becomes i, while i becomes -1, etc.:



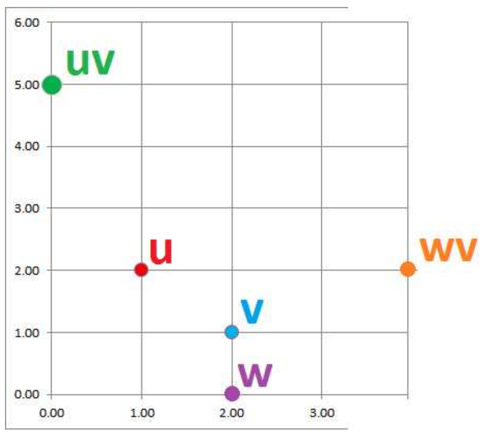
$1 \cdot i = i$	rotation from 0 degrees to 90
$i \cdot i = i^2 = -1$	rotation from 90 degrees to 180
$-1 \cdot i = -i$	rotation from 180 degrees to 270
$-i \cdot i = -i^2 = 1$	rotation from 270 degrees to 360
and so on.	

Example 4.10.2: complex multiplication

A more complex example:

$$\begin{array}{ll} u & = 1 + 2i \\ v & = 2 + i \\ \hline uv & = 2 + 4i + i + 2i^2 \\ & = (2-2) + (4+1)i \\ & = 0 + 5i \end{array}$$

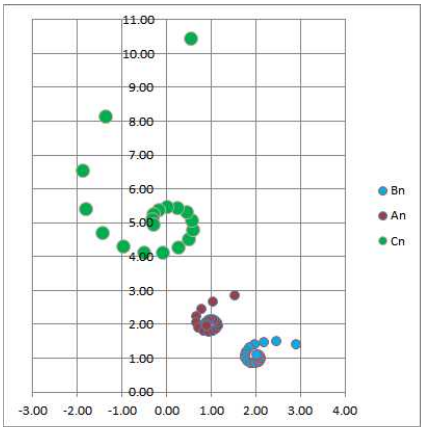
The rotation of v is visible:



In contrast, we can see the result of multiplying v by $w = 2$: no rotation.

So, the imaginary part of c is responsible for rotation.

How does multiplication affect topology?



Theorem 4.10.3: Product Rule for Complex Sequences

If sequences z_k, u_k *converge, then so does* $z_k \cdot u_k$*, and*

$$\lim_{k\to\infty} (z_k \cdot u_k) = \lim_{k\to\infty} z_k \cdot \lim_{k\to\infty} u_k\,.$$

Proof.

Suppose

$$z_k = a_k + b_k i \to a + bi \quad \text{and} \quad u_k = p_k + q_k i = p + qi\,.$$

Then, according to the *Componentwise Convergence Theorem* above, we have:

$$a_k \to a,\ b_k \to b \quad \text{and} \quad p_k \to p,\ q_k \to q\,.$$

Then, by the *Product Rule for numerical sequences*, we have:

$$a_k p_k \to ap,\ a_k q_k \to aq,\ b_k p_k \to bp,\ b_k q_k \to bq\,.$$

Then, as we know,

$$z_k \cdot u_k = (a_k p_k - b_k q_k) + (a_k q_k + b_k p_k)i \to (ap - bq) + (aq + bq)i = (a + bi)(p + qi)\,,$$

by the *Sum Rule for numerical sequences.*

Theorem 4.10.4: Quotient Rule for Complex Sequences

If sequences $z_k,\ u_k$ *converge (with* $u_k \neq 0$*), then so does* z_k/u_k*, and*

$$\lim_{k\to\infty} \frac{z_k}{u_k} = \frac{\lim_{k\to\infty} z_k}{\lim_{k\to\infty} u_k}\,,$$

provided

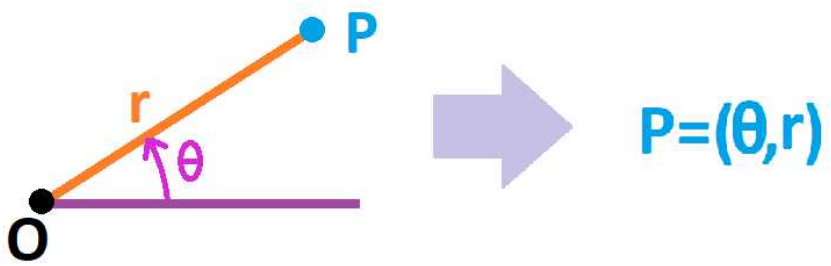
$$\lim_{k\to\infty} u_k \neq 0\,.$$

Just like real numbers!

Exercise 4.10.5

Prove the last theorem.

In addition to the standard, Cartesian, representation, a complex number $x = a + bi$ can be defined in terms of the *polar coordinates.*



We just continue our correspondence with a new one:

$$\boxed{a + bi \ \longleftrightarrow\ (a, b) \ \longleftrightarrow\ (\theta, r)}$$

The two quantities θ and r become the following:

Definition 4.10.6: modulus and argument

Suppose z is a complex number.

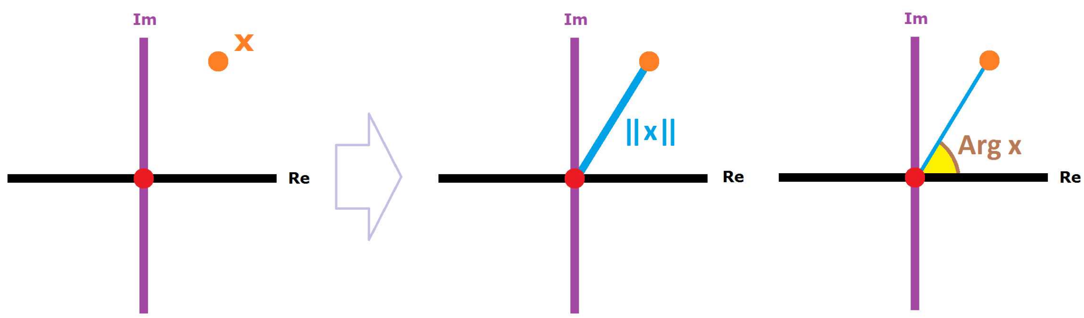
1. The distance from the location of z on the complex plane to the origin O

is called the *modulus* of z denoted by:

$$\boxed{||z||}$$

2. The angle of the line through the point z from the origin O with the x-axis is called the *argument* of z denoted by:

$$\boxed{\operatorname{Arg}(z)}$$



A simple examination tells us how to transition between the two coordinate systems:

Theorem 4.10.7: Conversion of Complex Numbers

Suppose $x = a + bi$ is a complex number. Then, we have:

1. *The modulus of z is found by:*

$$\boxed{||z|| = \sqrt{a^2 + b^2}}$$

2. *The argument of z is found by:*

$$\boxed{\operatorname{Arg}(z) = \arctan \frac{b}{a}}$$

Any two real numbers $r \geq 0$ and $0 \leq \theta < 2\pi$ can serve as those. It is called the *geometric representation* of the complex number:

$$\boxed{z = r\big[\cos\theta + i\sin\theta\big]}$$

The *algebra* takes a new form too. We don't need the new representation to compute addition and multiplication by real numbers, but we *need* it for multiplication.

What is the product of two complex numbers:

$$z_1 = r_1\big[\cos\varphi_1 + i\sin\varphi_1\big] \quad \text{and} \quad z_2 = r_2\big[\cos\varphi_2 + i\sin\varphi_2\big]\ ?$$

Consider:

$$\begin{aligned} z_1 z_2 &= r_1\big[\cos\varphi_1 + i\sin\varphi_1)\big] \cdot r_2\big[\cos\varphi_2 + i\sin\varphi_2\big] \\ &= r_1 r_2\big[\cos\varphi_1 + i\sin\varphi_1\big] \cdot \big[\cos\varphi_2 + i\sin\varphi_2\big] \\ &= r_1 r_2\big(\cos\varphi_1\cos\varphi_2 + i\sin\varphi_1\cos\varphi_2 + \cos\varphi_1\sin\varphi_2 + i^2\sin\varphi_1\sin\varphi_2\big)\,. \end{aligned}$$

We utilize the following trigonometric identities (Volume 1):

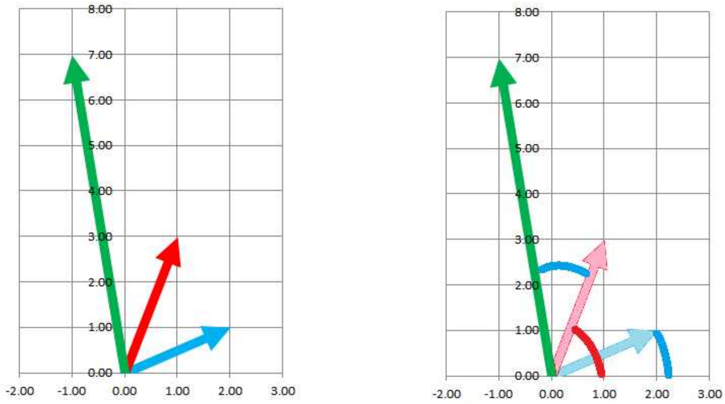
$$\cos a\cos b - \sin a\sin b = \cos(a+b) \quad \text{and} \quad \cos a\sin b + \sin a\cos b = \sin(a+b)\,.$$

Then,

$$z_1 z_2 = r_1 r_2 \big[\cos(\varphi_1 + \varphi_2) + i \sin(\varphi_1 + \varphi_2) \big] \,.$$

Example 4.10.8: geometric representation of multiplication

We can see the above computation on the complex plane:



We have proven the following:

Theorem 4.10.9: Multiplication of Complex Numbers

When two complex numbers are multiplied, their moduli are multiplied and the arguments are added.

In other words, we have:

$$\boxed{\begin{aligned} & r_1 \big[\cos\varphi_1 + i \sin\varphi_1) \big] \cdot r_2 \big[\cos\varphi_2 + i \sin\varphi_2 \big] \\ = \; & r_1 r_2 \big[\cos(\varphi_1 + \varphi_2) + i \sin(\varphi_1 + \varphi_2) \big] \end{aligned}}$$

Exercise 4.10.10

Suppose we have a convergent sequence of complex numbers. Consider the sequence of the moduli and the sequence of the arguments of the terms of the sequence and prove that they converge.

Exercise 4.10.11

(a) Represent the following complex number in the standard form: $(2+3i)(-1+2i)$. Indicate the real and imaginary parts. (b) Find its modulus and argument.

Exercise 4.10.12

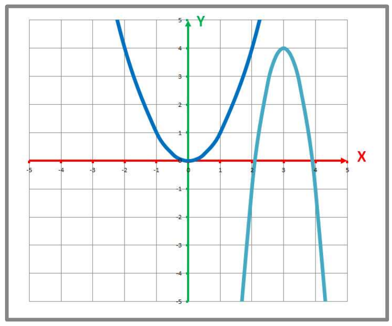
Simplify $(1+i)^2$.

Exercise 4.10.13

(a) Find the roots of the polynomial $x^2 + 2x + 2$. (b) Find its x-intercepts. (c) Find its factors.

Exercise 4.10.14

What can you say about the imaginary parts of the roots of these quadratic polynomials?



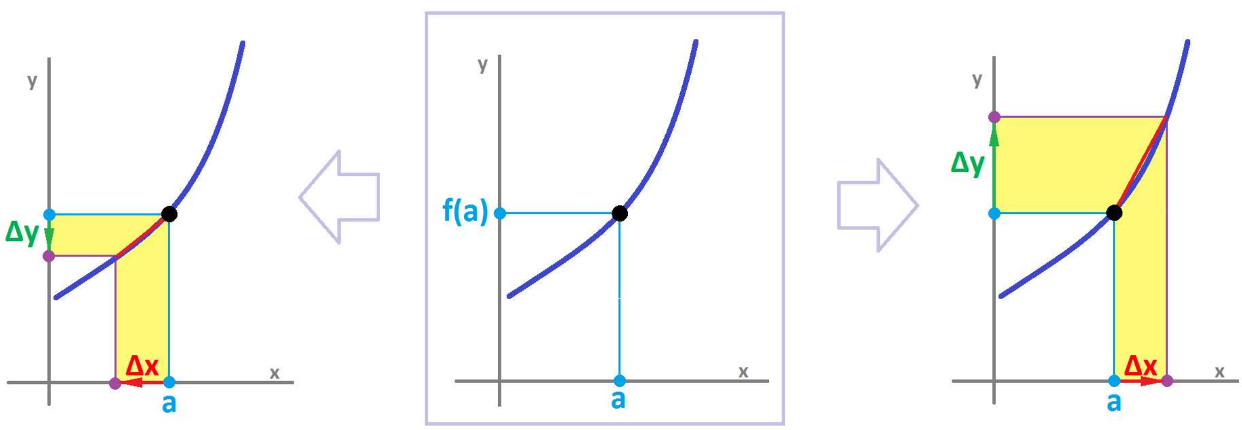
4.11. Discrete forms

In this section, we outline an alternative view on discrete calculus: differences, difference quotients, sums, and Riemann sums. However, we will deal with something even more fundamental than those four: We move beyond *functions* defined on nodes of augmented partitions.

Suppose we have a function $y = f(x)$ and we are to study its behavior around a point $x = a$. The difference quotient of f at a is the following:

$$\left.\frac{\Delta f}{\Delta x}\right|_{x=a} = \text{ the slope of the secant line through } (a, f(a)) = \frac{\text{rise of the secant line}}{\text{run of the secant line}}$$

We can always see Δx, Δy on the graph:



Thus, we have:

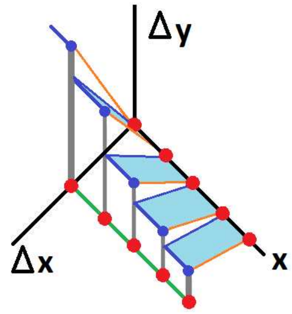
- Δx is the run of the secant line.
- Δy is the rise of the secant line.

There is a functional relation between them:

- y depends on x via
$$y = f(x)\,.$$
- Δy depends on Δx *and* x via
$$\Delta y = \frac{\Delta f}{\Delta x} \cdot \Delta x\,.$$

The latter, trivial, equation refers to a specific location, $x = a$ and $y = f(a)$, on the xy-plane, and it is a relation between the two *new variables* as the old ones have been specified.

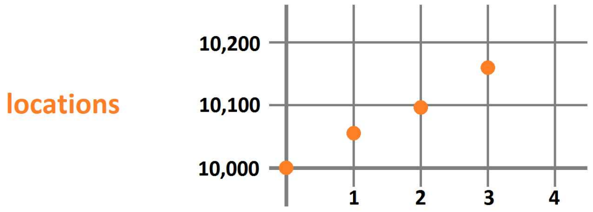
We can, furthermore, make these variables explicit:



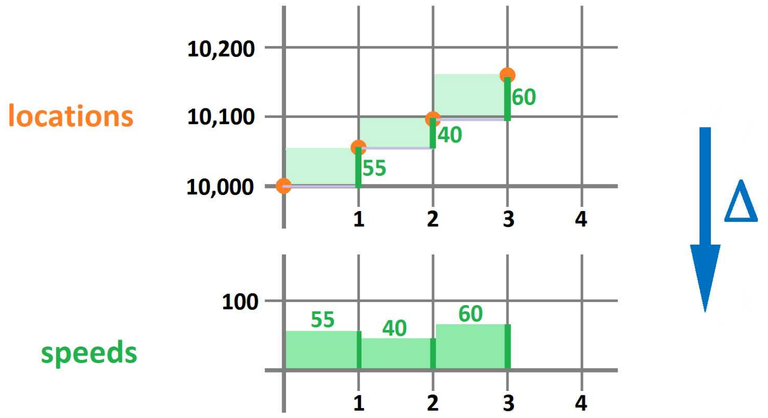
Below, we adopt a simpler, if less explicit, approach.

Example 4.11.1: broken speedometer

Recall what we started with in the very beginning. Suppose the speedometer is broken and in order to estimate how fast we are driving, we look at the odometer every hour:



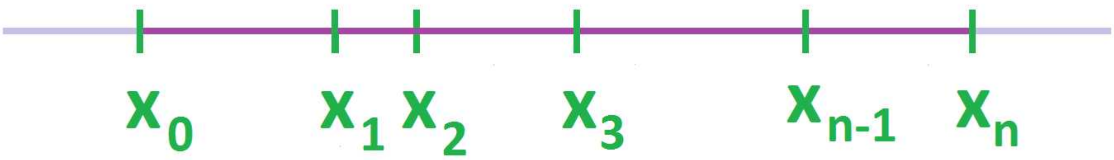
That's a discrete 0-form. To find the displacement for every hour, we just look at the differences:



That's a discrete 1-form. Alternatively, the odometer is broken and we look at the speedometer to sample the velocity and then, via the Riemann sums, find the displacement.

Let's start over.

Suppose we have a partition of some interval $[a, b]$ on the x-axis.



This time, we won't add secondary nodes but, instead, consider a *cell decomposition* of the interval.

There are two types of pieces in the interval:

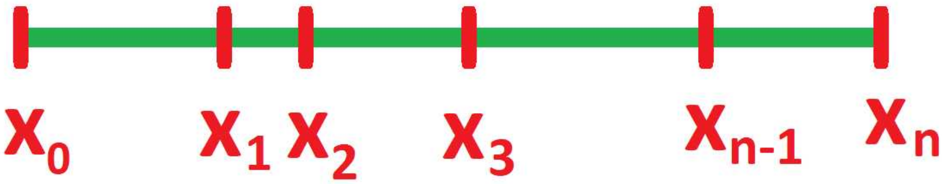
- the *nodes*:

$$x = x_k,\ k = 0, 1, ..., n$$

- the *edges*:

$$c_k = [x_{k-1}, x_k],\ k = 1, ..., n$$

The increments of x are still $\Delta x_k = x_k - x_{k-1}$.



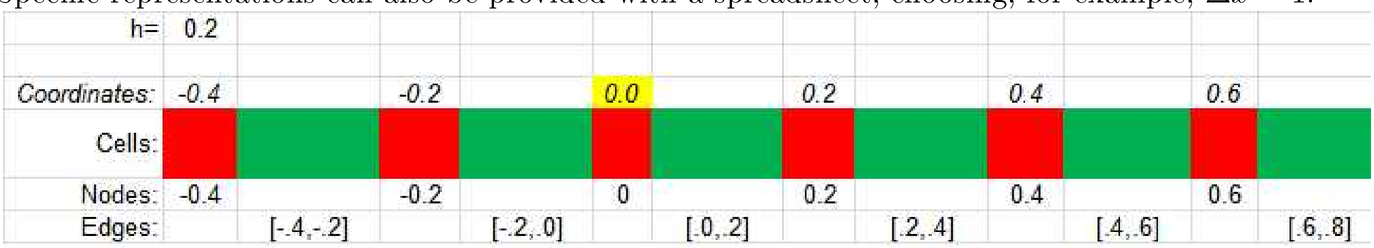
We introduce these names:

Definition 4.11.2: cells

- The nodes are called 0-*cells*.
- The edges are called 1-*cells*.

Example 4.11.3: cell decomposition

Specific representations can also be provided with a spreadsheet, choosing, for example, $\Delta x = 1$:



h=	0.2											
Coordinates:	*-0.4*		*-0.2*		*0.0*		*0.2*		*0.4*		*0.6*	
Cells:												
Nodes:	-0.4		-0.2		0		0.2		0.4		0.6	
Edges:		[-.4,-.2]		[-.2,.0]		[.0,.2]		[.2,.4]		[.4,.6]		[.6,.8]

You can see how every other cell is a square and every other is stretched horizontally to emphasize the different nature of these cells: nodes vs. edges.

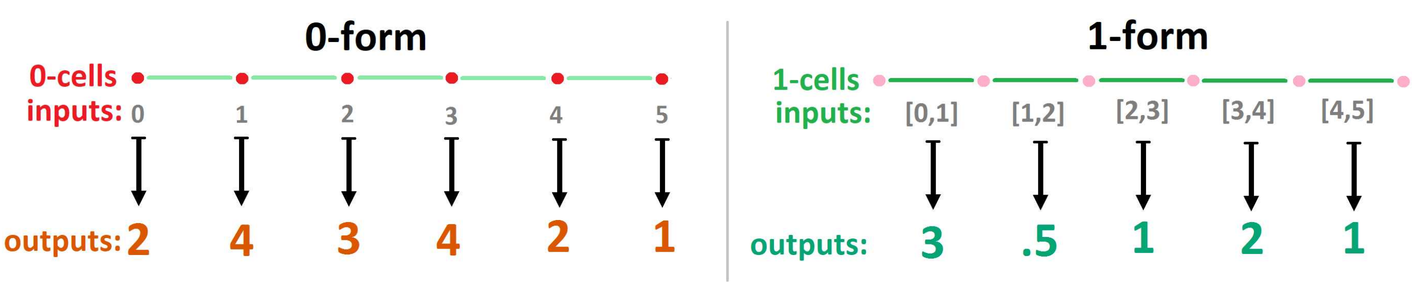
In the motion interpretation, there is a number (the location) associated with each node and a number (the displacement) associated with each edge.

Definition 4.11.4: discrete form

For a given partition (of an interval or the whole real line), we define the following:

- A *discrete form of degree* 0 is a real-valued function with 0-cells (nodes) as inputs.
- A *discrete form of degree* 1 is a real-valued function with 1-cells (edges) as inputs.

We use arrows to picture these functions as correspondences:



Here we have two:

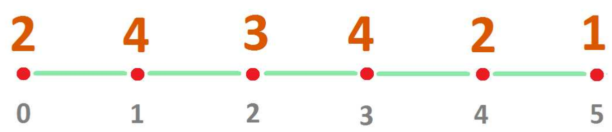
- a discrete 0-form

$$f : 0 \mapsto 2,\ 1 \mapsto 4,\ 2 \mapsto 3,\ \dots$$

- a discrete 1-form

$$s : [0,1] \mapsto 3,\ [1,2] \mapsto .5,\ [2,3] \mapsto 1,\ \dots$$

A more compact way to visualize is this:



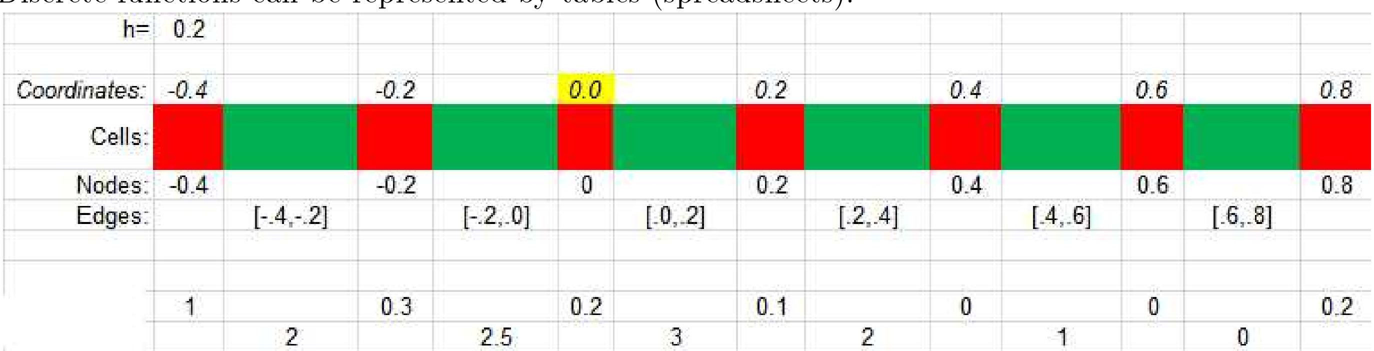
We can also *list the values* of the two functions:

- a discrete 0-form f:
$$f(0) = 2,\ f(1) = 4,\ f(2) = 3,\ \dots$$

- a discrete 1-form s:
$$s\Big([0,1]\Big) = 3,\ s\Big([1,2]\Big) = .5,\ s\Big([2,3]\Big) = 1,\ \dots$$

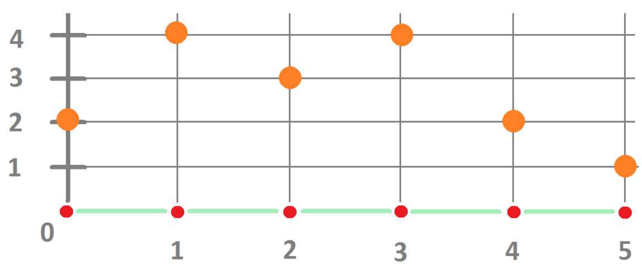
Example 4.11.5: forms with spreadsheet

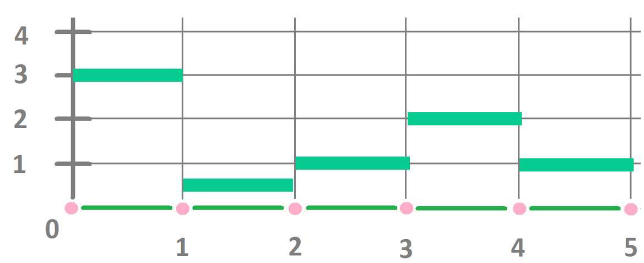
Discrete functions can be represented by tables (spreadsheets):



h=	0.2												
Coordinates:	*-0.4*		*-0.2*		*0.0*		*0.2*		*0.4*		*0.6*		*0.8*
Cells:													
Nodes:	-0.4		-0.2		0		0.2		0.4		0.6		0.8
Edges:		[-.4,-.2]		[-.2,.0]		[.0,.2]		[.2,.4]		[.4,.6]		[.6,.8]	
	1		0.3		0.2		0.1		0		0		0.2
		2		2.5		3		2		1		0	

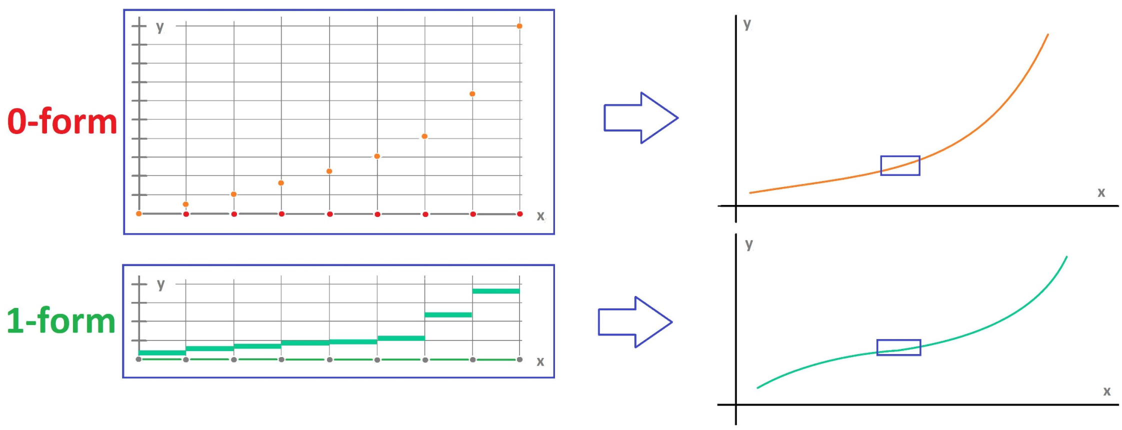
The most common way to visualize a function is with its *graph*, which consists of points on the xy-plane with $y = f(x)$:





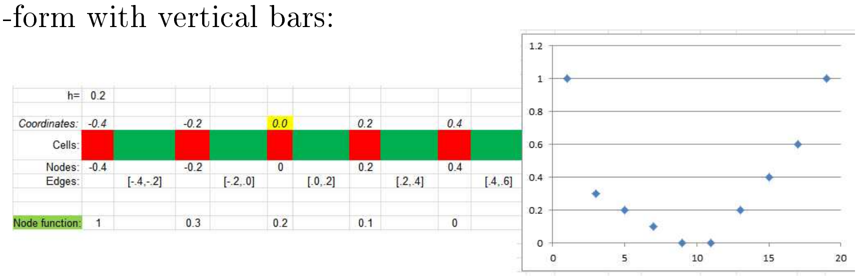
- For a discrete 0-form, x is a node, a number, and $y = f(x)$ is also a number. Together, they produce (x, y), a point on the xy-plane (with the x-axis split into cells as shown above).
- For a discrete 1-form, $[A, B]$ is an interval in the x-axis, and $y = g([A, B])$ is a number. Together, they produce a collection of points on the xy-plane such as (x, y) for every x in $[A, B]$. The result is a horizontal segment.

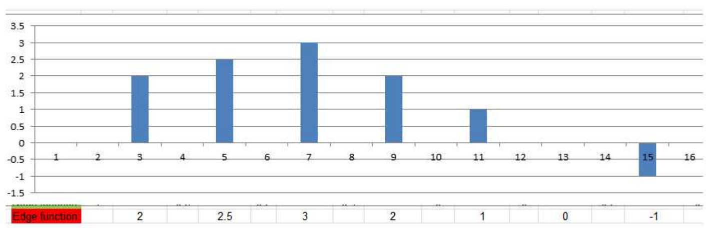
Even though these functions may consist of unrelated pieces, it is possible that we can see a *continuous curve* if we zoom out:



Example 4.11.6: graphs of forms with spreadsheet

To underscore the difference between the two, the graph of a discrete 0-form is shown with dots and that of a discrete 1-form with vertical bars:





Next, we discuss some of the *calculus* issues.

Example 4.11.7: difference

Let's consider an example of motion. Suppose a 0-form p gives the position of a person and suppose the following:

- At time n hours, we are at the 5-mile mark: $p(n) = 5$.
- At time $n+1$ hours, we are at the 7-mile mark: $p(n+1) = 7$.

We don't know what exactly has happened during this hour but the simplest assumption would be that we have been walking at a constant speed of 2 miles per hour.

Now, instead of our velocity function v assigning this value to each instant of time during this period, it is assigned to the *whole* interval:

$$v\Big|_{[n,n+1]} = 2\,,$$

or better:

$$v\Big([n, n+1]\Big) = 2\,.$$

This way, the elements of the domain of the velocity function are the edges and the resulting function is a discrete 1-form!

The functions, when defined on the nodes, change abruptly and, consequently, the change over every interval $[A, B]$ is simply the *difference of values* at the nodes, from right to left:

$$f(B) - f(A)\,.$$

The output of this simple computation is then assigned to the interval $[A, B]$:

$$\boxed{[A, B] \mapsto f(B) - f(A)}$$

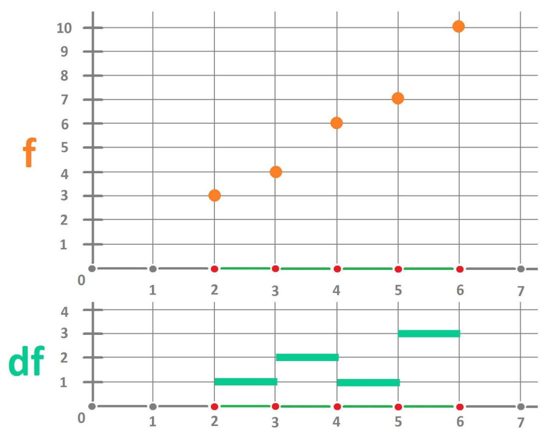
Just as before, the difference stands for the change of the function.

Definition 4.11.8: difference of discrete 0-form

The *difference of a discrete* 0*-form* f is a discrete 1-form given by its values at each edge:

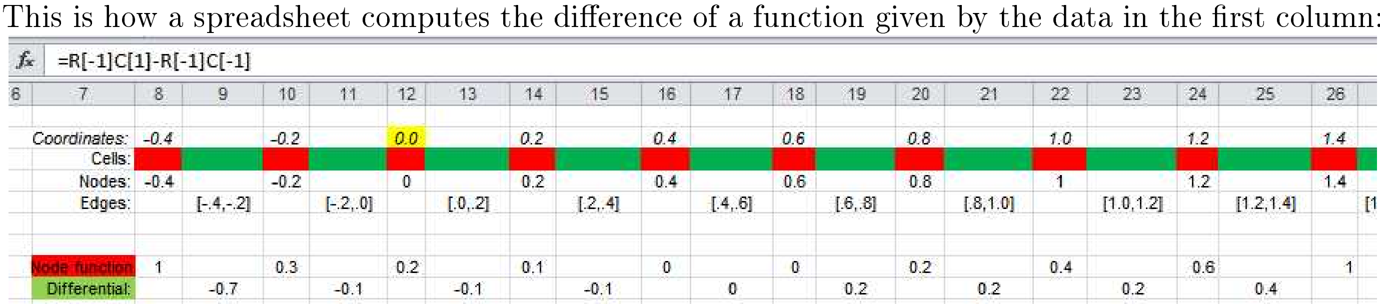
$$\boxed{\Delta f\,(c_k) = f(x_k) - f(x_{k-1})}$$

The relation between a 0-form and its difference is illustrated below:



Example 4.11.9: difference with spreadsheet

This is how a spreadsheet computes the difference of a function given by the data in the first column:



=R[-1]C[1]-R[-1]C[-1]

6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
	Coordinates:	*-0.4*		*-0.2*		*0.0*		*0.2*		*0.4*		*0.6*		*0.8*		*1.0*		*1.2*		*1.4*
	Cells:																			
	Nodes:	-0.4		-0.2		0		0.2		0.4		0.6		0.8		1		1.2		1.4
	Edges:		[-.4,-.2]		[-.2,.0]		[.0,.2]		[.2,.4]		[.4,.6]		[.6,.8]		[.8,1.0]		[1.0,1.2]		[1.2,1.4]	
	Node function	1		0.3		0.2		0.1		0		0		0.2		0.4		0.6		1
	Differential:		-0.7		-0.1		-0.1		-0.1		0		0.2		0.2		0.2		0.4	

Example 4.11.10: computing differences

When the discrete 0-forms are represented by formulas, the computations are straightforward ($h = 1$)

with a chance of simplification:

$$\begin{array}{llll} (1) & f(n)=3n^2+1 & \Longrightarrow & \Delta f\,(c_n)=(3n^2+1)-(3(n-1)^2+1)=6n-3 \\ (2) & g(n)=\dfrac{1}{n} & \Longrightarrow & \Delta g\,(c_n)=\dfrac{1}{n}-\dfrac{1}{n-1}=-\dfrac{1}{n(n-1)} \text{ for } n\neq 0,1 \\ (3) & p(n)=2^n & \Longrightarrow & \Delta p\,(c_n)=2^n-2^{n-1}=2^{n-1} \end{array}$$

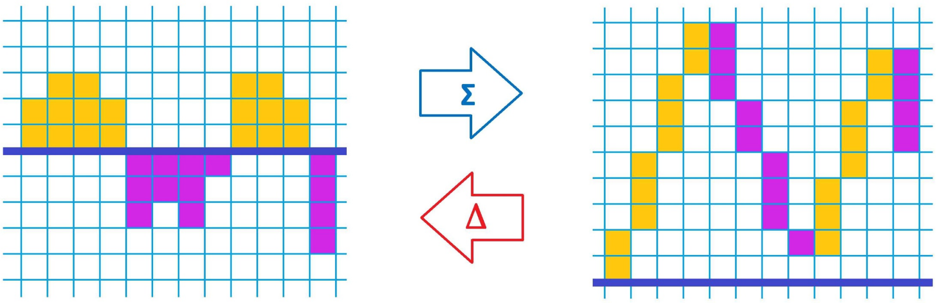
Definition 4.11.11: sum of discrete 1-form

The *sum of a discrete 1-form* g is a discrete 0-form given by its value at each node $x_k,\ 1\le k\le n$, of a partition of $[a,b]$ by:

$$\sum_{[a,x_k]} g = g(c_1)+g(c_2)+...+g(c_k)\,,$$

where $c_1,c_2,...,c_n$ are the edges of the partition.

The fundamental relation is between the differences and sums. The result is very similar to the one presented in Chapter 1; the two operations cancel each other in either order:



First, we have a 0-form and a 1-form:

- if f is defined at the nodes $x_k,\ k=0,1,2,...,n$, of the partition, then
- the difference g of f is defined at the edges of the partition by:

$$g(c_k)=f(x_k)-f(x_{k-1})\,.$$

Theorem 4.11.12: Fundamental Theorem of Discrete Calculus I

Suppose f is a discrete 0-form. Then, for each node x of the partition, we have:

$$\sum_{[a,x]}(\Delta f)=f(x)-f(a)\,.$$

Second, we have a 1-form and a 0-form:

- if g is defined at the edges $c_k,\ k=1,2,...,n$, of the partition, then
- the sum f of g is defined recursively at the nodes of the partition by:

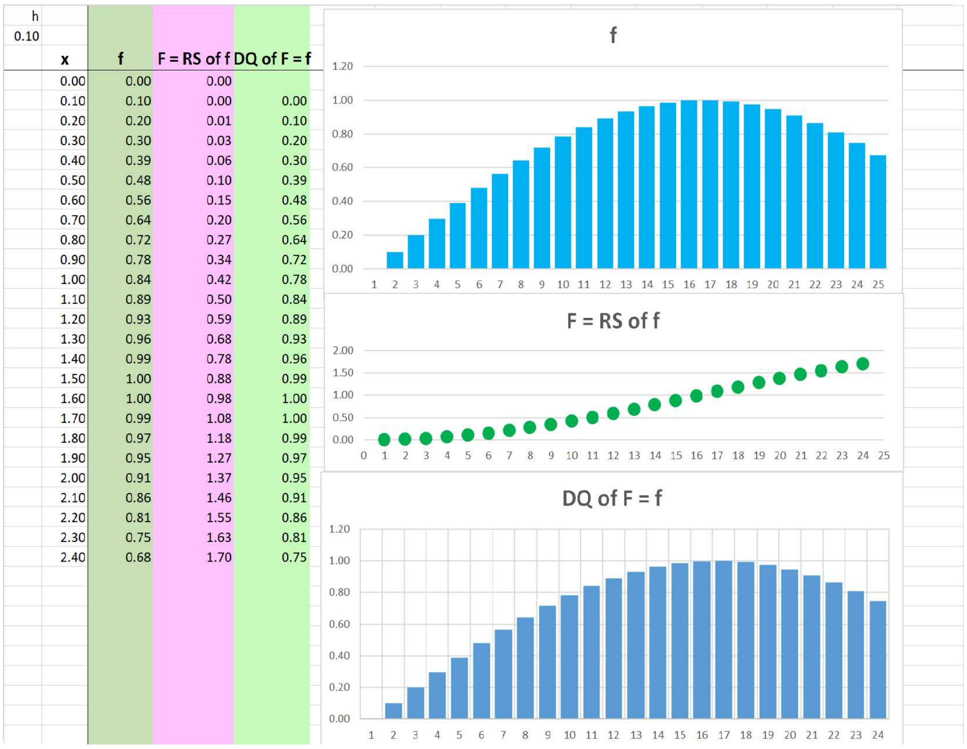
$$f(x_k)=f(x_{k-1})+g(c_k)\,.$$

Theorem 4.11.13: Fundamental Theorem of Discrete Calculus II

Suppose g is a discrete 1-form. Then, we have:

$$\Delta\left(\sum_{[a,x]} g\right) = g\,.$$

Just as in Chapter 1, we carry out the computations with a spreadsheet:



x	f	F = RS of f	DQ of F = f
0.00	0.00	0.00	
0.10	0.10	0.00	0.00
0.20	0.20	0.01	0.10
0.30	0.30	0.03	0.20
0.40	0.39	0.06	0.30
0.50	0.48	0.10	0.39
0.60	0.56	0.15	0.48
0.70	0.64	0.20	0.56
0.80	0.72	0.27	0.64
0.90	0.78	0.34	0.72
1.00	0.84	0.42	0.78
1.10	0.89	0.50	0.84
1.20	0.93	0.59	0.89
1.30	0.96	0.68	0.93
1.40	0.99	0.78	0.96
1.50	1.00	0.88	0.99
1.60	1.00	0.98	1.00
1.70	0.99	1.08	1.00
1.80	0.97	1.18	0.99
1.90	0.95	1.27	0.97
2.00	0.91	1.37	0.95
2.10	0.86	1.46	0.91
2.20	0.81	1.55	0.86
2.30	0.75	1.63	0.81
2.40	0.68	1.70	0.75

Next, compositions.

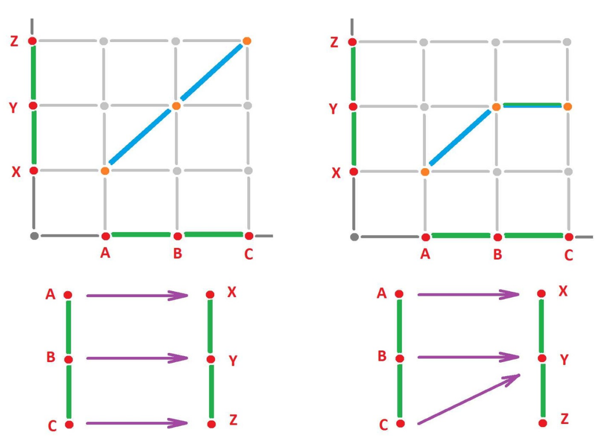
Next, there are no compositions of forms! For example, there is no way to execute these consecutively:

- 0-cell $\mapsto$ number, followed by
- 0-cell $\mapsto$ number

To be able to form a composition, one of these has to map cells to cells:

- 0-cell $\mapsto$ 0-cell $\mapsto$ number
- 1-cell $\mapsto$ 1-cell $\mapsto$ number

We create a composition $q \circ p$ of a 0- or 1-form q with another function or form p only when the values of p are 0- and 1-cells respectively. To define such a special function, we will require that the function p assigns a k- or $(k-1)$-cell to each k-cell:



There is also a special requirement:

Definition 4.11.14: cell function

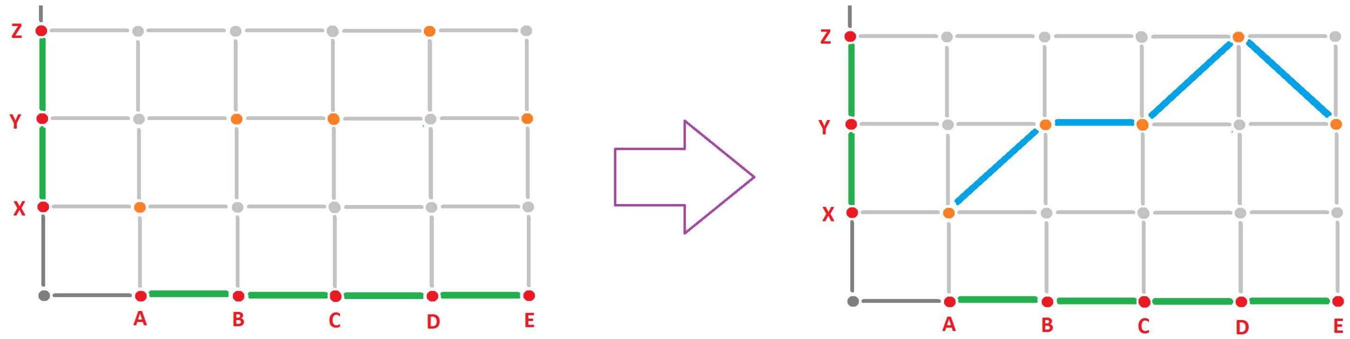
A *cell function* $y = p(x)$ is a function that assigns

- a node to each node, and
- an edge or a node to each edge,

in such a way that the end-points of each edge remain end-points:

$$\boxed{p\big([u,v]\big) = \big[p(u), p(v)\big]}$$

The requirement guarantees "continuity":



Because of the property, the values of a cell function on the edges can be reconstructed from its values on the nodes. The former is then analogous to the *difference* of the cell function.

For convenience, we assume that Δ is zero when computed over any node x.

Below is an analog of the chain rule:

Theorem 4.11.15: Chain Rule

The difference of the composition of two functions is the composition of the difference of the latter with the former; i.e., for any cell function $x = p(t)$ from $[a, b]$ to $[c, d]$ and any 0-form $y = g(x)$ on $[c, d]$, we have the differences satisfy:

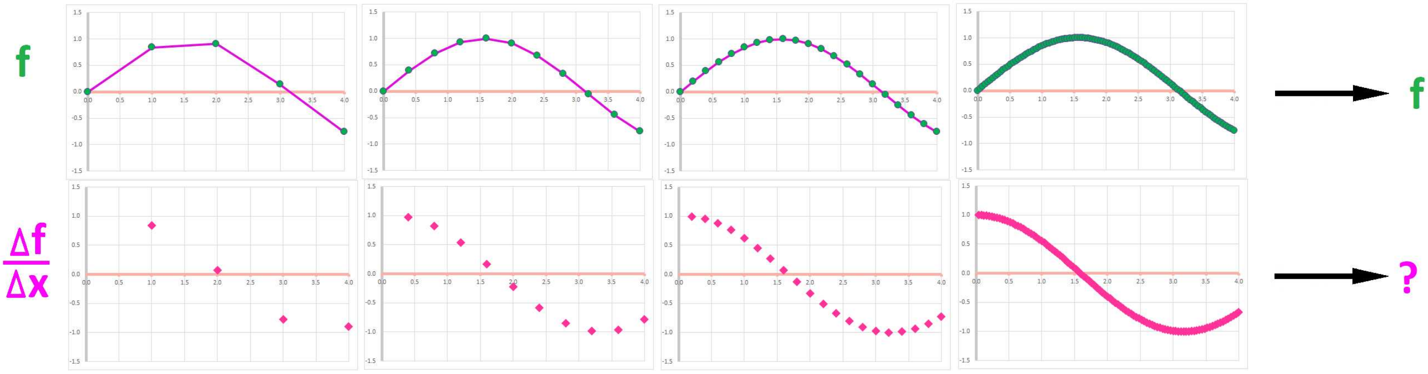
$$\Delta(g \circ p) = \Delta g \circ p\,.$$

In other words, we have for each edge s:

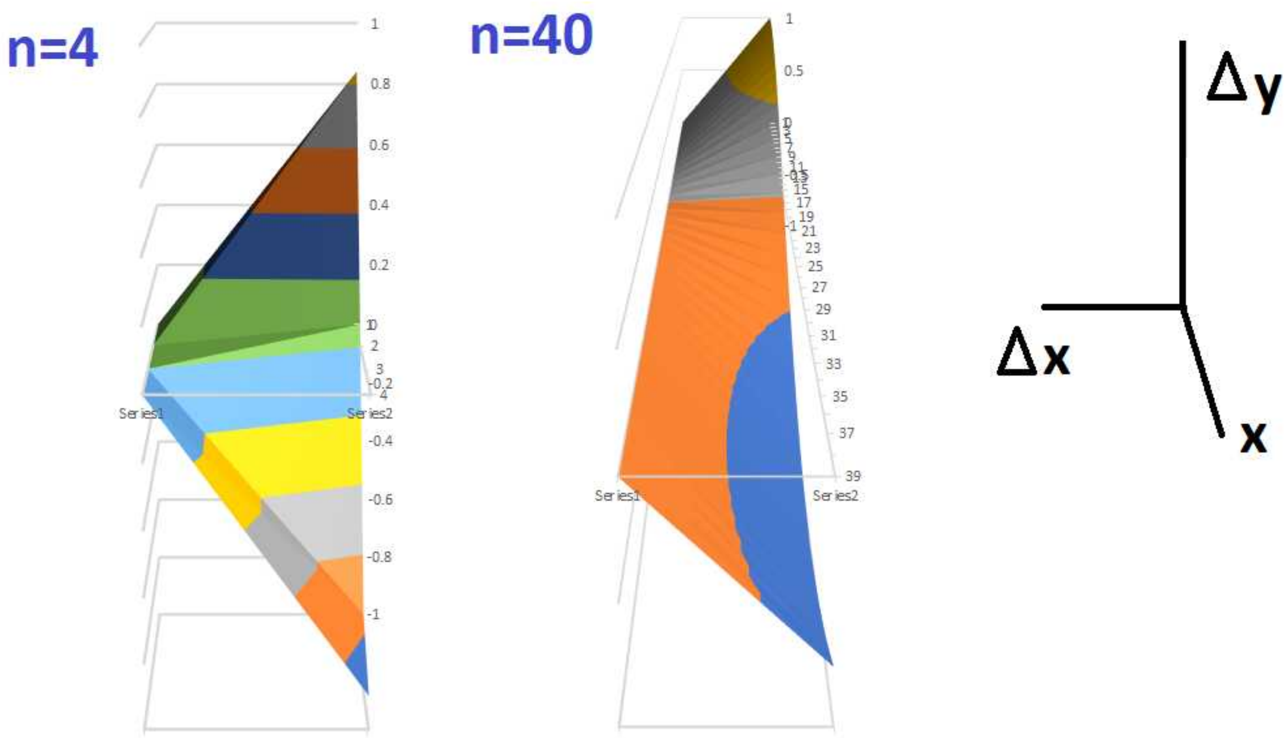
$$\Delta(g \circ p)(s) = \Delta g\,(p(s))\,.$$

Just as in the case of a traditional treatment, we ask: If we sample a function defined on an interval, what happens if we refine the partitions?

First, what happens to the *difference quotient* as $\Delta x \to 0$? It converges to the derivative:

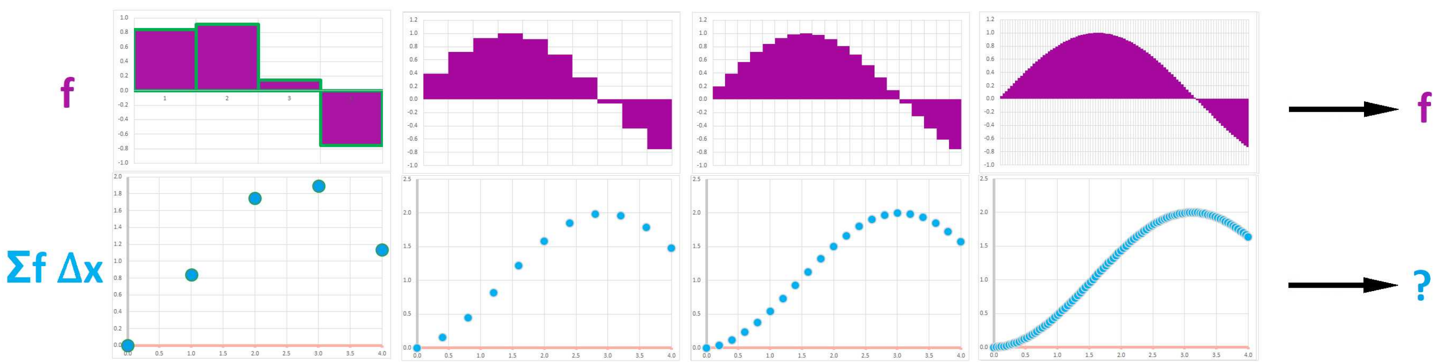


But what happens to the *difference* as $\Delta x \to 0$? A new concept emerges:



It is a *differential form of degree* 1 discussed in the next section.

Second, what happens to the *Riemann sum* as $\Delta x \to 0$? It converges to the Riemann integral:



But what happens to the *sum* as $\Delta x \to 0$? There is no convergence! Unless the function being sampled is a differential form of degree 1. No need for a new concept.

4.12. *Differential forms*

In this section, we outline an alternative view on calculus: derivative and Riemann integrals. However, we will deal with something even more fundamental than those two: We move beyond *functions* defined on intervals.

Question: Is the derivative $\dfrac{dy}{dx}$ a fraction?

The answer that followed the definition was an emphatic No!

A more advanced answer we give here is: Yes, here's how and why.

Suppose we have a function $y = f(x)$ and we are to study its behavior around a point $x = a$. The derivative

at a is the following:

$$\left.\frac{dy}{dx}\right|_{x=a} = \text{ the slope of the tangent line through } (a, f(a)) = \frac{\text{rise of the tangent line}}{\text{run of the tangent line}}$$

This *is* a fraction after all!

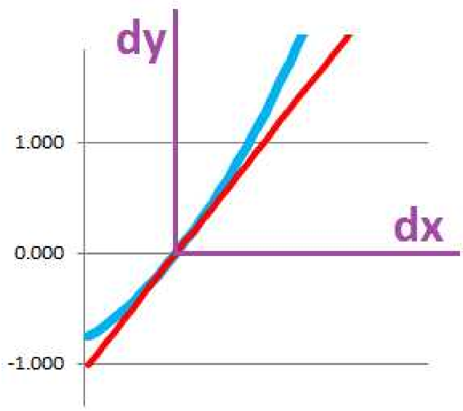
Example 4.12.1: $dx - dy$-plane

Specifically, suppose $f(x) = x^2 + 2x$. At $a = 0$, we have $f(0) = 0$, so our interest is the point $(0, 0)$. Then,

$$\left.\frac{dy}{dx}\right|_{x=0} = 2x + 2\Big|_{x=0} = 2\,.$$

If this is a fraction, what would be the meaning of this:

$$dy = 2 \cdot dx\,?$$



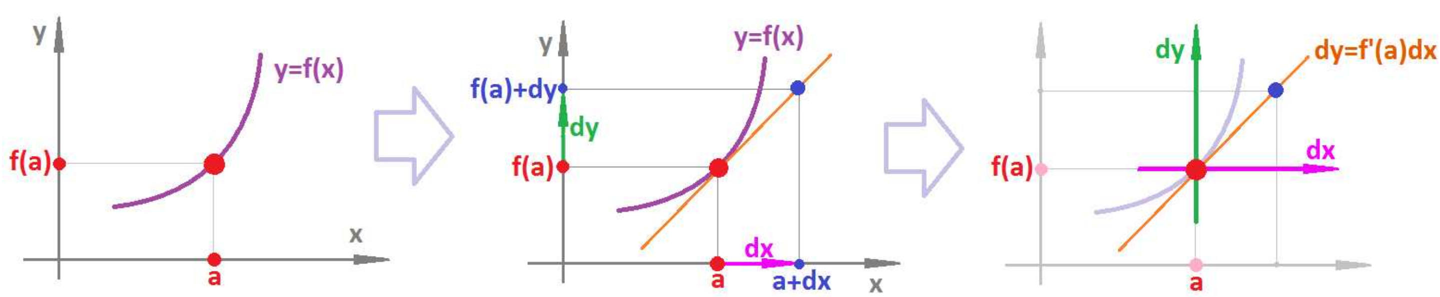
It is the equation of the tangent line written with respect to dy and dx.

Thus, the equation

$$dy = f'(a) \cdot dx$$

refers to a specific location, $x = a$ and $y = f(a)$, on the xy-plane and it is a relation between the two *new variables* as the old ones have been specified.

We can always see dx, dy on the graph:



Thus, we have:

- dx is the run of the tangent line.
- dy is the rise of the tangent line.

They are called the *differentials* of x and y respectively.

Warning!

Here, $X = dx$ and $Y = dy$ are just certain variables related to x and y respectively; the latter depends on the former linearly:

$$Y = m \cdot X\,.$$

The algebra may come from the example above:

- y depends on x via $y = f(x)$.
- dy depends on x and dx via $dy = f'(x)dx$.

Example 4.12.2: linearization

Given a function $f(x) = x^2$, find its best linear approximation at $a = 1$. Since $f'(x) = 2x$, we see that $f'(a) = f'(1) = 2$ and, therefore, the best linear approximation of f at $a = 1$ is

$$L(x) = f(a) + f'(a)(x - a) = 1 + 2(x - a)\,.$$

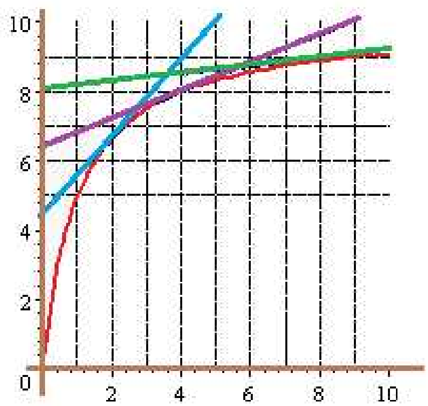
Now we re-interpret these quantities:
1. $dx = x - a$
2. $dy = L(x) - L(a)$

Then, we have:

$$dy = 2 \cdot dx\,.$$

The equation expresses our derivative in terms of these new variables, the differentials. We capture the relation between the increment of x and that of y – *close to* a. Indeed, y grows twice as fast as x. We acquire this information by introducing a new coordinate system (dy, dx). In this coordinate system, the best linear approximation (given by the tangent line) becomes simply a linear function.

The analysis presented above applies to every point – and to all points at once:



Recall also from Chapter 1 how we learned to look at the integral differently:

$$\int_a^b k(x)\,dx\,.$$

We change *what* we integrate. Instead of a function, $k(x)$, it is a *differential form*, $k(x) \cdot dx$.

Definition 4.12.3: differential form of degree 1

A *differential form of degree* 1, or simply a 1-form, is defined as a function of two variables:

$$\varphi = \varphi(x, dx) = k(x) \cdot dx\,,$$

where $y = k(x)$ is a function of x.

The function is simply linear with respect to the second variable.

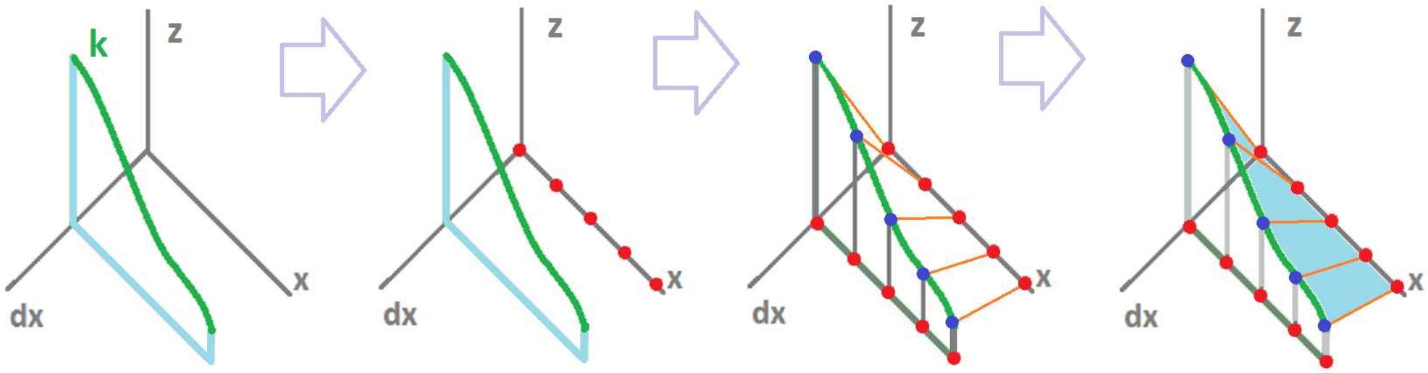
Warning!

The symbol "$\cdot$" stands for multiplication and it is often omitted.

Let's plot the graph of such a function below:

- First, we plot the graph of k (green) above the line $dx = 1$.
- Then, we observe that φ is 0 when $dx = 0$ and plot points on the x-axis (red).
- Finally, we connect these dots to the curve with *straight lines* (orange).

The result is this surface:



As presented above, a differential form may come from the following:

$$y = f(x) \text{ at } x = a \implies \frac{dy}{dx} = f'(a)\,,$$

and, furthermore,

$$\implies \quad dy = f'(a) \cdot dx\,.$$

A differential form is a relation between the two extra variables, once the relation between the old ones has been specified. The dependence between the differentials varies from location to location. So, dx is the differential of x, which is a variable separate from, but related to, x.

Recall also how the *Chain Rule*, in the Leibniz notation, was interpreted as a "cancellation" of du:

$$\frac{dy}{dx} = \frac{dy}{\cancel{du}}\frac{\cancel{du}}{dx}\,.$$

Now we can see that it is indeed a cancellation, when du is not zero.

Differential forms are what we integrate. For simplicity, instead of using partitions and discrete forms to define the integral, we just refer to the "usual" integral:

Definition 4.12.4: integral of 1-form

The *integral of a 1-form* $\varphi = k\,dx$ over an interval $[a, b]$ is defined to be

$$\int_{[a,b]} \varphi = \int_a^b k(x)\,dx\,.$$

Then the form $k\,dx$ is *integrable* whenever k is integrable.

Let's take a look at how we represented the formula of *integration by substitution* in Chapter 2:

$$\int f(u) \cdot \frac{du}{\cancel{dx}}\ \cancel{dx} = \int f(u)\,du\,.$$

In light of the new definition, this is a literal cancellation.

Exercise 4.12.5

Show that the sum, but not the product, of two 1-forms is also a 1-form. And so is a multiple of a differential form.

Differentiation is also translated into the language of differential forms.

Definition 4.12.6: differential form of degree 0

A *differential form of degree* 0, or simply a 0-form, is any function $y = f(x)$ of x.

Just as discrete forms, differential forms of different degrees are interconnected.

Definition 4.12.7: exterior derivative of differential form of degree 0

The *exterior derivative* df of a differential form f of degree 0 is defined to be the 1-form given by:

$$\boxed{df = f'(x)\, dx}$$

This notation is used along with others used when the name of the function that relates x and y is not provided:

Differential

$$df \quad dy \quad d(\ )$$

For example, we have as before:

$$y = x^2 \implies dy = 2x\, dx\,.$$

We can also have:

$$d(x^2) = 2x\, dx\,.$$

Thus, the exterior derivative of a function contains all information about its derivative, and vice versa. However, the former provides a direct answer to this question:

- ▶ If we are at $x = a$ and make a step dx, what is the step dy of y?

Example 4.12.8: displacement

Suppose x is time and $y = f(x)$ is the location at time x. The exterior derivative provides a direct answer to this question:

- ▶ Suppose x is time and $y = f(x)$ is the location at time x. If at time $x = a$ we are at $y = f(a)$ and then we move for a short short time dx, how far will we go?

It's the velocity multiplied by the time increment:

$$\text{Displacement } = f'(a) \cdot dx\,,$$

but only when the velocity, f', is constant. In the general case, this is an *estimate*.

Example 4.12.9: integration by substitution

We have used this algebra for *integration by substitution*. Consider the integral:

$$\int_0^2 2x \sin x^2\, dx\,.$$

The idea is to introduce a new variable

$$u = x^2 .$$

Here is a familiar computation interpreted in a new way:

$u = x^2 \implies du = 2x\,dx$	$du = 2x\,dx$ Exterior derivative
$\implies \displaystyle\int_{x=0}^{x=2} 2x \sin x^2\,dx$	$\displaystyle\int_{[0,2]} 2x \sin x^2\,dx$ $[0,2] \to [0^2, 2^2]$
$= \displaystyle\int_{u=0^2}^{u=2^2} \sin u\,du$	$= \displaystyle\int_{[0^2,2^2]} \sin u\,du$
$= -\cos u \Big\|_{u=0}^{u=4}$	$= -\cos u \Big\|_{u=0}^{u=4}$
$= -\cos 4 - (-\cos 0)$	$= -\cos 4 - (-\cos 0)$

Our definition of differential form treats the integrand as a simple case of *multiplication* of two numbers. That is why we are at liberty to algebraically manipulate these expressions the way we have.

The following is a simple re-statement with our new notation of a familiar theorem (Chapter 1):

Theorem 4.12.10: Fundamental Theorem of Calculus

Suppose φ *is a* 1*-form integrable on interval* $[a,b]$*. Then,*

$$\int_{[a,b]} \varphi = F(b) - F(a) ,$$

for any 0*-form* F *that satisfies:*

$$dF = \varphi .$$

In order to study a real-valued function $y = f(x)$ *and* its change, we now keep track of *two variables*:

- the locations, x vs. y, and
- the directions, dx vs. dy.

The relation is as follows:

$$\boxed{(x, dx) \mapsto (y, dy) = \big(f(x), f'(x)dx\big)}$$

What is the relation to *discrete forms*?

We know that discrete forms created by sampling a function and refining partitions converges to a differential form of the same degree. Can we reverse this correspondence?

Any function can be sampled. It matters where:

1. Sampling at the primary nodes produces a discrete 0-form.
2. Sampling at the secondary nodes produces a discrete 1-form.

A better idea, however, is to sample a differential form of the corresponding degree.

Theorem 4.12.11: Discrete and Differential Forms

1. *A differential* 0*-form evaluated at the nodes of the partition is a discrete* 0*-form.*

2. *A differential 1-form evaluated – by integration – at the intervals of the partition is a discrete 1-form; i.e., if φ is a differential 1-form, the corresponding discrete 1-form is defined by:*

$$\boxed{s\Big([A,B]\Big)=\int_{[A,B]}\varphi}$$

So, a differential form of degree 1 *is* its integrals:

Integral of differential form

$$\begin{array}{cccccc} & \text{“evaluate”} & & & & \\ & \downarrow & & & & \\ & \int & \underbrace{(3x^3+\sin x)\,dx} & = & 0 & \\ \text{input} \to & [-1,1] & \uparrow & & \uparrow & \\ & & \text{differential form} & & \text{output} & \end{array}$$

If we set the motion interpretation aside, the purpose of the new concept is to make a careful distinction between the location, x, and the direction, dx:

- ▶ How fast are we going from this location in that direction?

There is only one direction (and its opposite) for a function of one variable, but infinitely many for a function of two variables. We will see applications of differential forms in the multidimensional case in Volume 4.

Chapter 5: Series

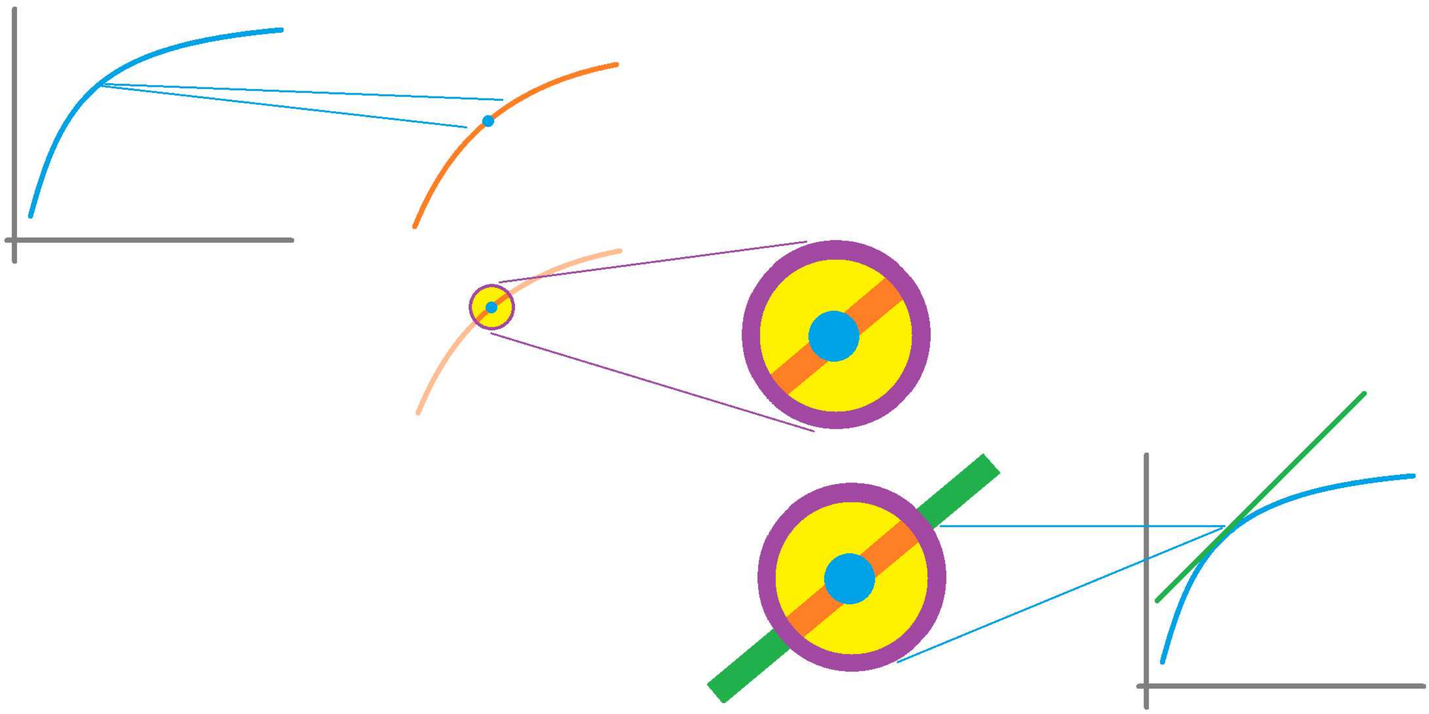
Contents

5.1. From linear to quadratic approximations

Approximating functions is like approximating numbers – such as π, e, or the Riemann integral of any function – but harder.

Recall from Volume 2 (Chapter 2DC-6) the meaning of *linearization*: We replace a given function $y = f(x)$ with a linear function $y = L(x)$ that best approximates it at a given point.

This function is called the best *linear* approximation of the function, and it happens to be the linear function the graph of which is the tangent line at the point. The replacement is justified by the fact that when you zoom in on the point, the tangent line will merge with the graph:



However, there is a more basic approximation: a *constant* function, $y = C(x)$.

Example 5.1.1: square root

Let's review this example from Chapter 2DC-6:

▶ How do we compute $\sqrt{4.1}$ without actually evaluating $f(x) = \sqrt{x}$?

We approximate. Specifically, in order to approximate the number of $\sqrt{4.1}$, we approximate the function $f(x) = \sqrt{x}$ "around" $a = 4$.

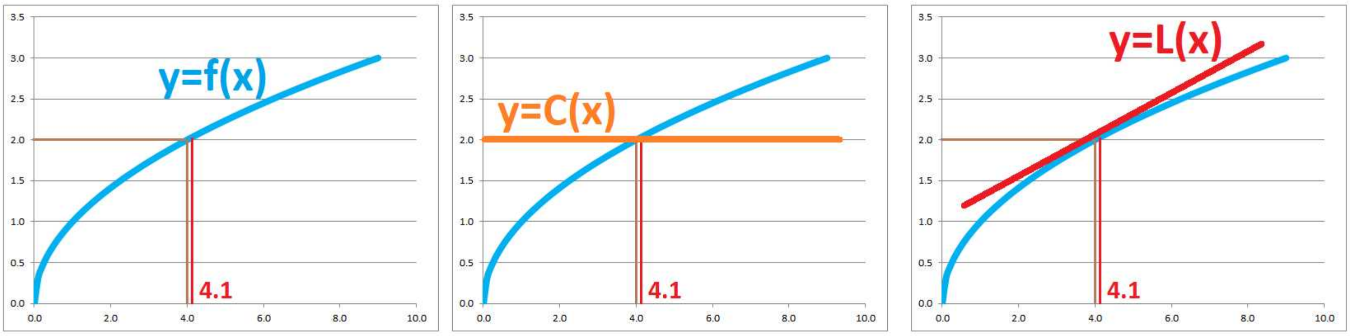
We first approximate the function with a *constant* function:

$$C(x) = 2\,.$$

This value is chosen because $f(a) = \sqrt{4} = 2$. Then we have:

$$\sqrt{4.1} = f(4.1) \approx C(4.1) = 2\,.$$

It is a crude approximation:



The other, linear, approximation is visibly better. We approximate the function with a *linear* function:

$$L(x) = 2 + \frac{1}{4}(x-4)\,.$$

This value is chosen because $f(a) = \sqrt{4} = 2$ and $f'(a) = \frac{1}{4}$. Then we have:

$$\sqrt{4.1} = f(4.1) \approx L(4.1) = 2 + \frac{1}{4}(4.1-4) = 2.025\,.$$

We have for a function $y = f(x)$ and a point $x = a$:

$$\begin{array}{ll} \text{The best constant approximation:} & C(x) = f(a)\,. \\ \text{The best linear approximation:} & L(x) = f(a) \quad + f'(a)(x-a)\,. \end{array}$$

We should notice early on that the latter just adds a new (linear) term to the former!

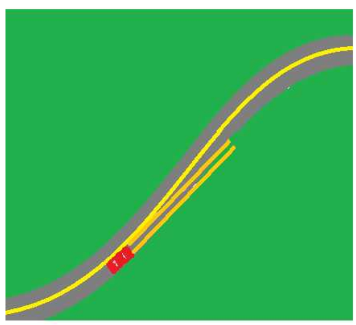
Warning!

The latter is better than the former – but only when we need more accuracy. Otherwise, the latter is worse because it requires more computation.

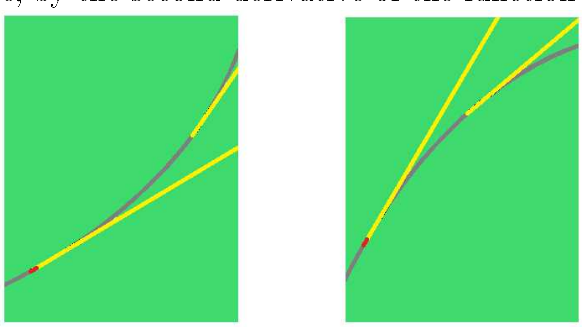
Can we do better than the *best* linear approximation? No. Can we do better than the best *linear* approximation? Yes.

Example 5.1.2: road curvature

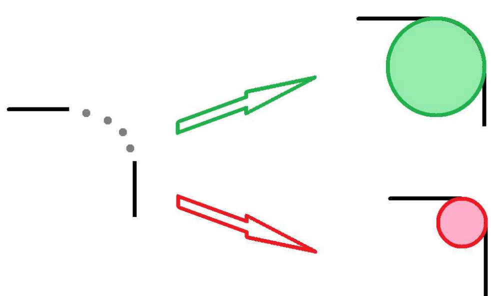
One can understand linearization as ignoring the shape of the road and concentrating on the headlights of the car (Chapter 2DC-3) one location at a time:



In Chapter 2DC-4, we also learned that the *curvature* of the road is determined by how fast the headlights turn and, therefore, by the second derivative of the function it represents:



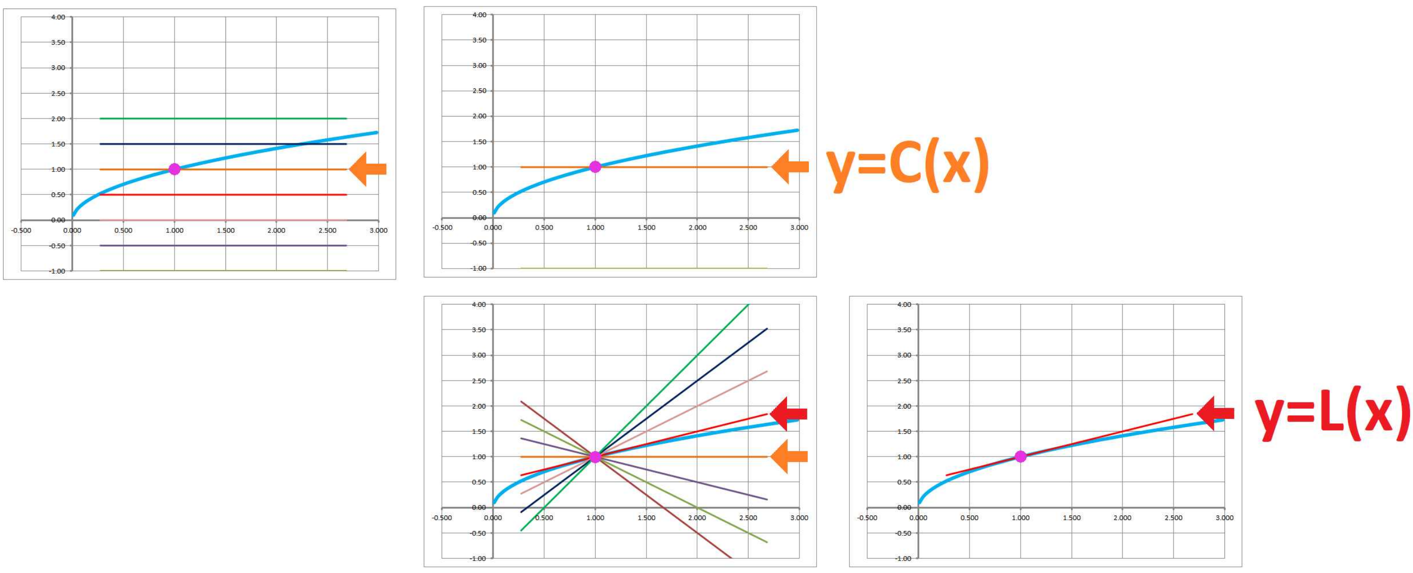
A further idea is to approximate the curve of the road with a *circle* of an appropriate radius – larger for lower curvature and smaller for higher curvature:



Every circle is a *quadratic* curve and, therefore, is seen as a *quadratic* approximation. Since the only thing we want from it is its curvature, we can replace the circle with another quadratic but simpler (and explicitly given) curve – the *parabola*. The curvature is further studied in Volume 4 (Chapter 4HD-2).

Below we illustrate how we attempt to approximate a function around the point $(1, 1)$ with *constant* functions

first; from those, we choose the horizontal line through the point. This line then becomes one of the many *linear* approximations of the curve that pass through the point; from those, we choose the tangent line:



Now, we shall see that these are just the two first steps in a *sequence* of approximations!

Indeed, the tangent line will become one of the many *quadratic* curves – parabolas – that pass through the point... and are tangent to the curve. Which one of *those* do we choose?

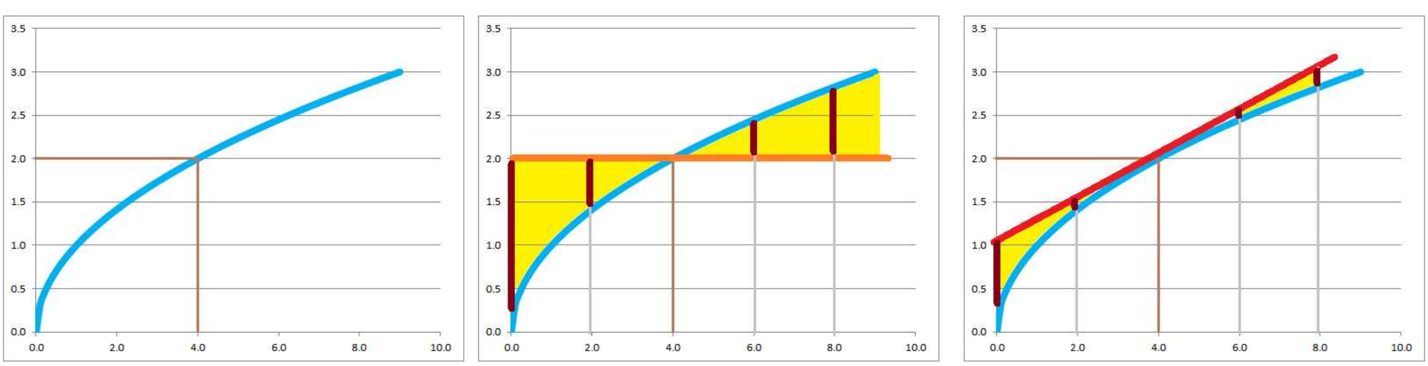
In order to answer that, we need to review and understand how the best constant and the best linear approximations were chosen. In what way are they the best?

Suppose a function $y = f(x)$ is given and we would like to approximate its behavior in the vicinity of a point, $x = a$, with another function $y = T(x)$. The latter is to be taken from some class of functions that we find suitable. A class of relatively simple functions that are also quite versatile is the *polynomials*.

What we need to consider is the *error*, i.e., the difference between the function f and its approximation T:

$$\boxed{E(x) = |f(x) - T(x)|}$$

The the error function is seen below:



We are supposed to minimize the error function in some way.

Of course, the error function $y = E(x)$ is likely to grow with no limit as we move away from our point of interest, $x = a$... but we don't care. We want to minimize the difference in the *vicinity* of a, which means making sure that the limit of the error as $x \to a$ goes to 0!

Theorem 5.1.3: Best constant approximation

Suppose f is continuous at $x = a$ and

$$C(x) = k$$

is any of its constant approximations (i.e., arbitrary constant functions). Then,

the error E of the approximation approaches 0 at $x = a$ if and only if the constant is equal to the value of the function f at $x = a$.

In other words, we have:

$$\lim_{x\to a}(f(x) - C(x)) = 0 \iff k = f(a)\,.$$

Proof.

Use the *Sum Rule* for limits and then the continuity of f:

$$0 = \lim_{x\to a}(f(x) - C(x)) = \lim_{x\to a} f(x) - \lim_{x\to a} C(x) = f(a) - k\,.$$

That's the analog of the following theorem from Volume 2 (Chapter 2DC-6):

Theorem 5.1.4: Best linear approximation

Suppose f is differentiable at $x = a$ and

$$L(x) = f(a) + m(x - a)$$

is any of its linear approximations. Then, the error E of the approximation approaches 0 at $x = a$ faster than $x - a$ if and only if the coefficient of the linear term is equal to the value of the derivative of the function f at $x = a$.

In other words, we have:

$$\lim_{x\to a}\frac{f(x) - L(x)}{x - a} = 0 \iff m = f'(a)\,.$$

Proof.

Use the *Sum Rule* for limits and then the differentiability of f:

$$0 = \lim_{x\to a}\frac{f(x) - L(x)}{x - a} = \lim_{x\to a}\frac{f(x) - f(a)}{x - a} - \lim_{x\to a} m = f'(a) - m\,.$$

Let's compare the conditions in the two theorems:

$$f(x) - C(x) \to 0 \quad\text{and}\quad \frac{f(x) - L(x)}{x - a} \to 0$$

The comparison reveals the similarity and the difference in how we minimize the error! The difference is in the degree: *how fast the error function goes to zero*. Indeed, we learned in Chapter 2DC-6 that the latter condition means that $f(x) - L(x)$ converges to 0 faster than $x - a$, i.e.,

$$f(x) - L(x) = o(x - a);$$

there is no such restriction for the former.

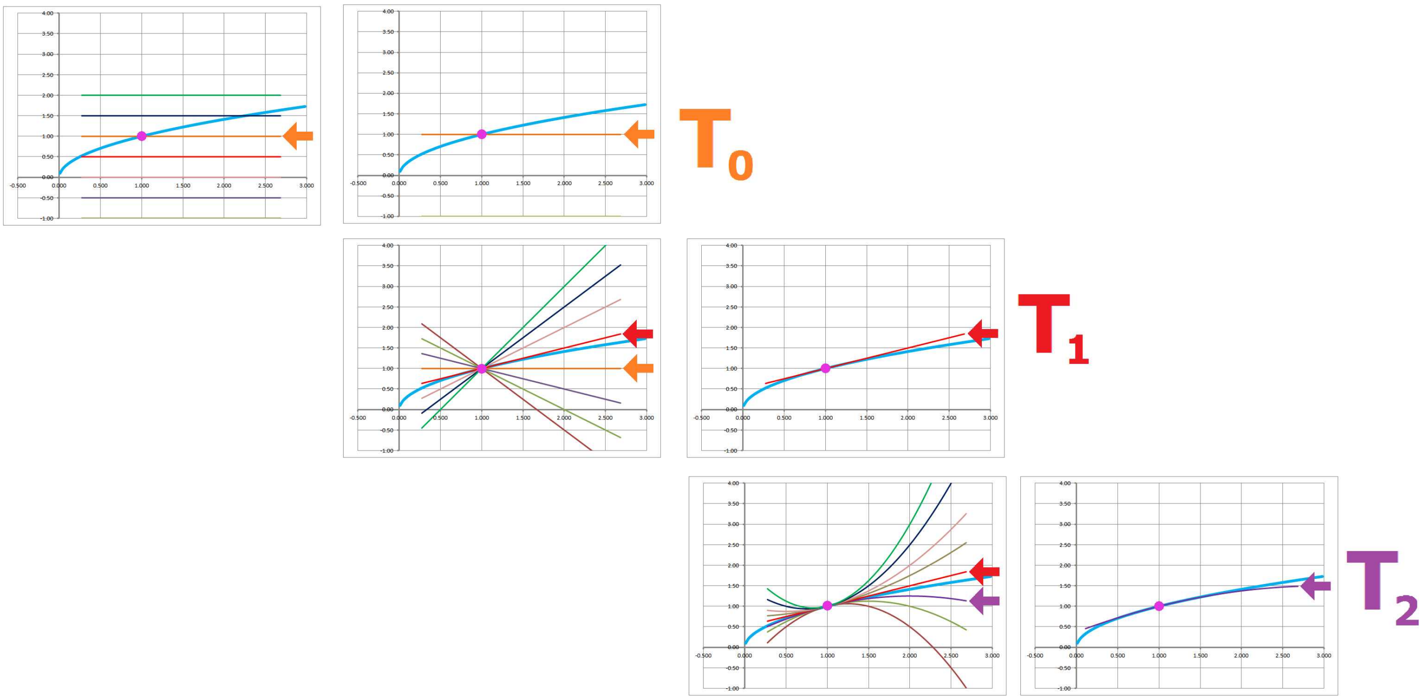
So far, this is what we have discovered:

▶ Linear approximations are built from the best constant approximation by adding a linear term.

The best one of those has the slope (its own derivative) equal to the derivative of f at a. How the sequence of approximations will progress is now clearer:

▶ Quadratic approximations are built from the best linear approximation by adding a quadratic term.

One of them might be the best:



To decide which one of those is the best, we think by analogy and try to make the error function E go to 0 but even faster:

Theorem 5.1.5: Best quadratic approximation

Suppose f is twice continuously differentiable at $x = a$ and

$$Q(x) = f(a) + f'(a)(x-a) + p(x-a)^2$$

is any of its quadratic approximations. Then, the error E of the approximation approaches 0 at $x = a$ faster than $(x-a)^2$ if and only if the coefficient of the quadratic term is equal to half of the value of the second derivative of the function f at $x = a$.

In other words, we have:

$$\lim_{x\to a}\frac{f(x)-Q(x)}{(x-a)^2} = 0 \iff p = \frac{1}{2}f''(a)\,.$$

Proof.

We apply *L'Hopital's rule* twice:

$$\begin{aligned}
0 &= \lim_{x\to a}\frac{f(x)-Q(x)}{(x-a)^2} && \text{First.}\\
&= \lim_{x\to a}\frac{f'(x)-f'(a)-2p(x-a)}{2(x-a)}\\
&= \lim_{x\to a}\frac{f''(x)-2p}{2} && \text{And second.}\\
&= \lim_{x\to a}\frac{f''(x)}{2} - p\\
&= \frac{1}{2}f''(a) - p\,.
\end{aligned}$$

Once again, the condition of the theorem means that $f(x) - Q(x)$ converges to 0 faster than $(x-a)^2$, or

$$f(x) - Q(x) = o((x-a)^2)\,.$$

We start to see a pattern:

- The degrees of the approximating polynomials are growing.
- The degrees of the derivatives being taken into account are growing too.

Example 5.1.6: square root

For the original example of $f(x) = \sqrt{x}$ at $a = 4$, we have:

$$\begin{array}{llll} f(x) & = \sqrt{x} & \Longrightarrow\ f(4) & = 2 \\ f'(x) & = (x^{1/2})' = \frac{1}{2}x^{-1/2} & \Longrightarrow\ f'(4) & = \frac{1}{4} \\ f''(x) & = \left(\frac{1}{2}x^{-1/2}\right)' = -\frac{1}{4}x^{-3/2} & \Longrightarrow\ f''(4) & = -\frac{1}{32} \end{array}$$

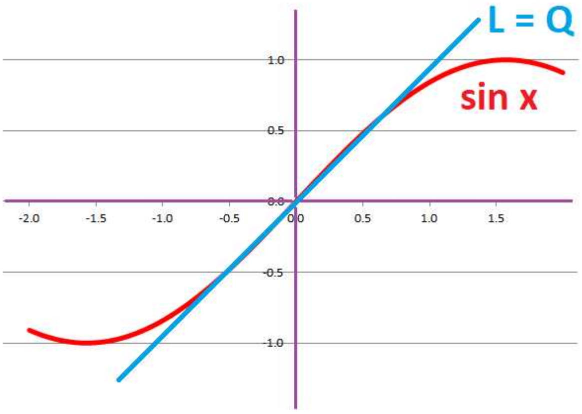
Therefore, we have:

approximations:				why best:
$C(x) =$	2			Same value as f.
$L(x) =$	2	$+\frac{1}{4}(x-4)$		...And same slope as f.
$Q(x) =$	2	$+\frac{1}{4}(x-4)$	$-\frac{1}{2\cdot 32}(x-4)^2$	...And same concavity as f.

Example 5.1.7: sine

Let's approximate $f(x) = \sin x$ at $x = 0$. First, the values of the function and the derivatives:

$$\begin{array}{llllll} f(x) & = \sin x & \Longrightarrow\ f(0) & = 0 & \Longrightarrow\ C(x) & = 0 \\ f'(x) & = \cos x & \Longrightarrow\ f'(0) & = 1 & \Longrightarrow\ L(x) & = x \\ f''(x) & = -\sin x & \Longrightarrow\ f''(0) & = 0 & \Longrightarrow\ Q(x) & = ? \end{array}$$



Therefore, the best quadratic approximation is:

$$Q(x) = 0 + 1(x-0) - \frac{0}{2}(x-0)^2 = x\,.$$

Same as the linear! Why? Because the sine is odd.

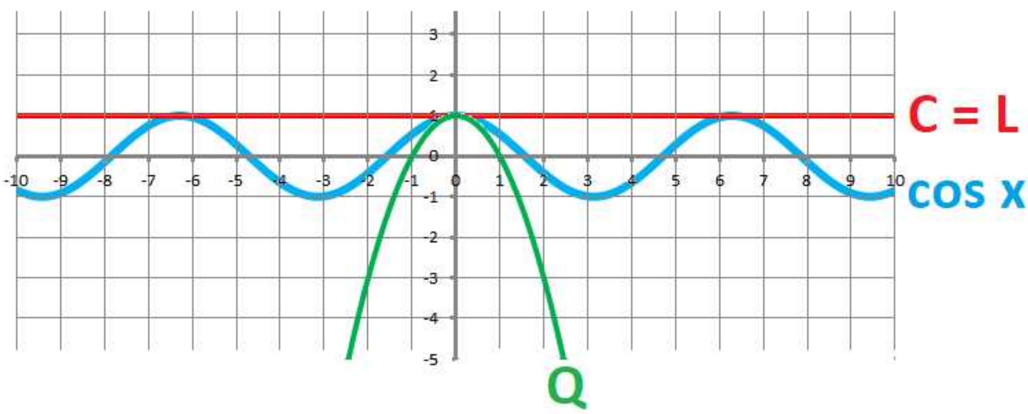
Example 5.1.8: cosine

Let's approximate $f(x)=\cos x$ at $x=0$. First, the values of the function and the derivatives:

$$\begin{array}{llllll} f(x) & =\cos x & \Longrightarrow & f(0) & =1 & \Longrightarrow & C(x) & =1 \\ f'(x) & =-\sin x & \Longrightarrow & f'(0) & =0 & \Longrightarrow & L(x) & =1 \\ f''(x) & =-\cos x & \Longrightarrow & f''(0) & =-1 & \Longrightarrow & Q(x) & =? \end{array}$$

Therefore, the best quadratic approximation is:

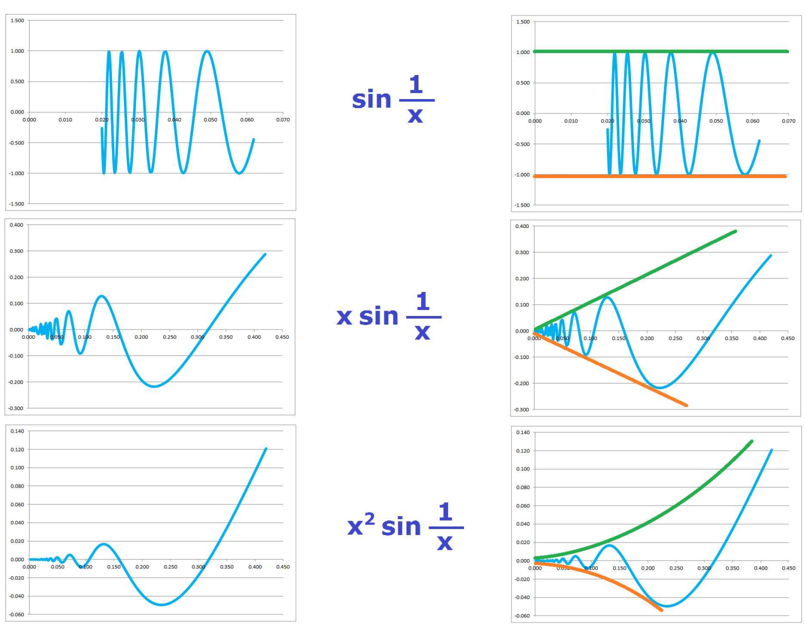
$$Q(x)=1+0(x-0)-\frac{1}{2}(x-0)^2=1-\frac{1}{2}x^2\,.$$



No linear term! Why? Because the cosine is even.

Example 5.1.9: edge behavior

The applicability of the theorems depends on the nature of the function. Consider these three familiar functions (Volume 2):



These are the results of our analysis:

1. Function $f(x)=\sin\frac{1}{x}$ (with $f(0)=0$) is not continuous at 0, and none of the theorems apply. There is no good approximation at 0, of any kind.
2. Function $g(x)=x\sin\frac{1}{x}$ (with $f(0)=0$) is continuous at 0, and the first theorem applies. But as it's not differentiable, and the other two theorems do not apply. The best constant approximation at 0 is $C(x)=0$, but it's not, and there is none, a good linear approximation.
3. Function $h(x)=x^2\sin\frac{1}{x}$ (with $f(0)=0$) is differentiable at 0, and the first two theorem apply. But it's not twice differentiable, and the last theorem does not apply. The best linear (and

constant) approximation at 0 is $L(x) = 0$ but it's not, and there is none, a good quadratic approximation.

We used sequences of *numbers* to approximate other numbers in Volume 2; now we will use sequences of *functions* to approximate other functions. In order to be able to go beyond quadratic in our sequence of polynomial approximations, we rename them according to their *degrees*:

$$\begin{array}{|ll|}\hline T_0(x) & = C(x) \\ T_1(x) & = L(x) \\ T_2(x) & = Q(x) \\ ... & \end{array}$$

Example 5.1.10: square root

Back to the original example of $f(x) = \sqrt{x}$ at $a = 4$. One can guess where this is going:

$$\begin{array}{rrrrrrrrrl|l}
\text{constant:} & & & & & & & 2 & = T_0(x) & f - T_0 = o(1) \\
\text{linear:} & & & & & \frac{1}{4}(x-4) & + & 2 & = T_1(x) & f - T_1 = o(x-a) \\
\text{quadratic:} & & & -\frac{1}{2\cdot 32}(x-4)^2 & + & \frac{1}{4}(x-4) & + & 2 & = T_2(x) & f - T_2 = o((x-a)^2) \\
\text{cubic:} & (?)(x-4)^3 & - & \frac{1}{2\cdot 32}(x-4)^2 & + & \frac{1}{4}(x-4) & + & 2 & = T_3(x) & f - T_3 = o((x-a)^3) \\
 & \vdots & & \vdots & & \vdots & & \vdots & \vdots & \vdots
\end{array}$$

We add a term every time we move down to the next degree; it's the recursive *sum* (as discussed in Chapter 1) of the sequence of the terms that keep appearing. The resulting sequence is called a "series".

Next, the general theory.

5.2. The Taylor polynomials

A polynomial is typically written in its *standard form*:

$$P(x) = a_0 + a_1x + a_2x^2 + ... + a_nx^n \,.$$

It's the sum of multiples (a linear combination) of the *powers* of x.

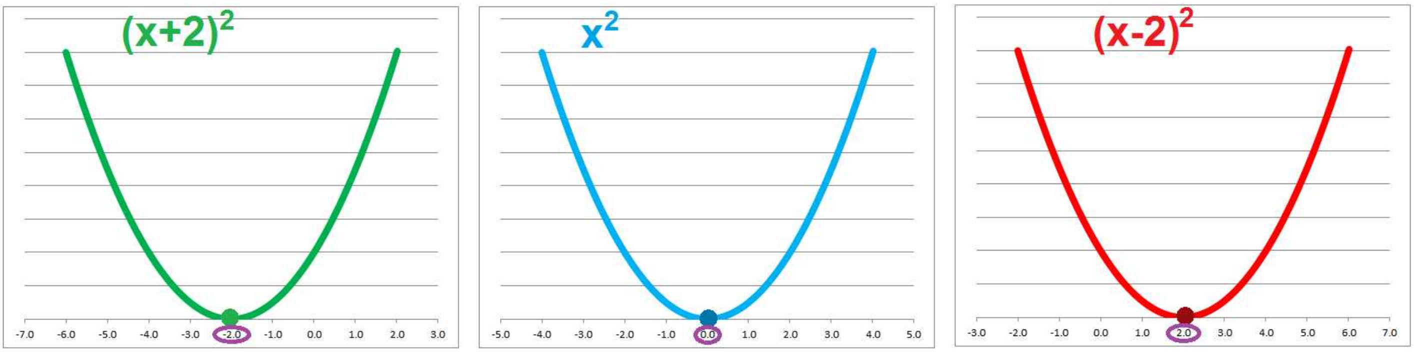
However, in order to be able to concentrate on a particular value of x, i.e., $x = a$, we want to express the polynomial in terms of the deviation of x from a, i.e., $x - a$. We find a special analog of the standard form of the polynomial; a polynomial is still the sum of powers, just not of x but of $(x - a)$.

A familiar example of such a transition comes from linear polynomials and their *point-slope form*:

$$L(x) = mx + b = m(x - a) + d \,.$$

There is also the *vertex form* of quadratic polynomials (Chapter 1PC-4):

$$Q(x) = a(x - h)^2 + k \,.$$



This is the general result:

Theorem 5.2.1: Centered Form of Polynomial

For each real number a, every degree n polynomial P can be represented in the form centered at $x = a$, i.e.,

$$P(x) = c_0 + c_1(x-a) + c_2(x-a)^2 + ... + c_n(x-a)^n ,$$

for some real numbers $c_0, ..., c_n$.

Proof.

We use this *change of variables*: $x \mapsto x - a$, a shift to the right by a.

The standard form is then just the form centered at $x = 0$.

It is from among these polynomials we will choose the best approximation at $x = a$.

Below is a table that shows the progress of better and better *polynomial approximations* following the ideas developed in the last section. The degrees of these polynomials are listed in the first column, while the first row shows the degree of the terms of these polynomials:

degrees	n		...		3		2		1		0	
0											c_0	$= T_0$
1									$c_1(x-a)$	$+$	c_0	$= T_1$
2							$c_2(x-a)^2$	$+$	$c_1(x-a)$	$+$	c_0	$= T_2$
3					$c_3(x-a)^3$	$+$	$c_2(x-a)^2$	$+$	$c_1(x-a)$	$+$	c_0	$= T_3$
$\vdots$					$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$
n	$c_n(x-a)^n$	$+$	...	$+$	$c_3(x-a)^3$	$+$	$c_2(x-a)^2$	$+$	$c_1(x-a)$	$+$	c_0	$= T_n$

Once again, the sequence T_n is the sum of the sequence $c_n(x-a)^n$.

How do we choose the best? We require that the error function converges to zero faster and faster:

Definition 5.2.2: Taylor polynomial

Suppose f is n times continuously differentiable at $x = a$. Then the *nth Taylor polynomial*, $n = 0, 1, 2, ...$, is defined recursively by:

$$T_0 = f(a), \quad T_n(x) = T_{n-1} + c_n(x-a)^n ,$$

under the requirement that the error of the approximation approaches 0 at $x = a$ faster than $(x-a)^n$:

$$\boxed{\lim_{x \to a} \frac{f(x) - T_n(x)}{(x-a)^n} = 0}$$

The coefficients $c_0, c_1, ..., c_n$ are called the *Taylor coefficients* of f.

In other words, the error is:

$$\boxed{E = f - T_n = o\big((x-a)^n\big)}$$

Applying the *Power Formula* of differentiation repeatedly produces more and more but smaller and smaller coefficients until it reaches 1:

$$x' = 1,\ (x^2)'' = (2x)' = 2,\ (x^3)'' = (3x^2)'' = (6x)' = 6,\ \dots$$

This is the general result.

Theorem 5.2.3: nth derivative of nth power

For any $n = 1, 2, 3...$, we have:

$$\boxed{(x^n)^{(n)} = n!}$$

Proof.

$$\begin{array}{llllllll}
 & (x^n)' & & & = & n\cdot & x^{n-1} & \Longrightarrow \\
\Longrightarrow & (x^n)'' & = & n\cdot & \left(x^{n-1}\right)' & = & n\cdot(n-1)\cdot & x^{n-2} & \Longrightarrow \\
\Longrightarrow & (x^n)''' & = & n\cdot(n-1)\cdot & \left(x^{n-2}\right)' & = & n\cdot(n-1)\cdot(n-2)\cdot & x^{n-3} & \Longrightarrow \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\Longrightarrow & (x^n)^{(n-1)} & = & n\cdot(n-1)\cdot\ldots\cdot 3\cdot & \left(x^2\right)' & = & n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot 3\cdot 2\cdot & x^1 & \Longrightarrow \\
\Longrightarrow & (x^n)^{(n)} & = & n\cdot(n-1)\cdot\ldots\cdot 3\cdot 2\cdot & x' & = & n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot 3\cdot 2\cdot & 1 & = n!
\end{array}$$

Same for powers of $(x-a)$:

$$\boxed{\big((x-a)^n\big)^{(n)} = n!}$$

So, that's why the *factorial* appears in the forthcoming formulas...

Exercise 5.2.4

Prove the last formula.

We jump straight to the answer. This is our main interest.

Theorem 5.2.5: Taylor Coefficients

Suppose f is n times continuously differentiable at $x = a$. Then the Taylor coefficients of f must be:

$$\boxed{c_0 = f(a),\ c_1 = f'(a),\ c_2 = \frac{1}{2}f^{(2)}(a),\ \dots,\ c_n = \frac{1}{n!}f^{(n)}(a)}$$

Proof.

From the last theorem, it follows that:

$$c_0 = T_0(a),\ c_1 = T_1'(a),\ c_2 = \frac{1}{2}T_2^{(2)}(a),\ ...,\ c_n = \frac{1}{n!}T_n^{(n)}(a)\,.$$

We use the limit in the definition for each $k = 0, 1, 2, ..., n$, as follows. Start with:

$$0:\quad 0 = \lim_{x\to a}\frac{f(x)-T_0(x)}{(x-a)^0} = \lim_{x\to a}(f(x)-T_0(x)) \implies T_0(a) = \lim_{x\to a}T_0(x) = \lim_{x\to a}f(x) = f(a)$$

We use the fact that both T_0 and f are continuous at $x = a$. It follows that $c_0 = T_0(a) = f(a)$. Next, we use *L'Hopital's Rule*:

$$1:\quad 0 = \lim_{x\to a}\frac{f(x)-T_1(x)}{(x-a)^1} = \lim_{x\to a}\frac{f'(x)-T_1'(x)}{1} \implies T_1'(a) = \lim_{x\to a}T_1'(x) = \lim_{x\to a}f'(x) = f'(a)$$

We use the fact that both T_1' and f' are continuous at $x = a$. Next, we apply L'Hopital's Rule twice:

$$2:\quad 0 = \lim_{x\to a}\frac{f(x)-T_2(x)}{(x-a)^2} = \lim_{x\to a}\frac{f''(x)-T''_2(x)}{2} \implies T''_2(a) = \lim_{x\to a}T''_2(x) = \lim_{x\to a}f''(x) = f''(a)$$

We use the fact that both T''_2 and f'' are continuous at $x = a$. And so on. For the last step, we apply L'Hopital's Rule n times:

$$n:\quad 0 = \lim_{x\to a}\frac{f(x)-T_n(x)}{(x-a)^n} = \lim_{x\to a}\frac{f^{(n)}(x)-T_n^{(n)}(x)}{n!} \implies T_n^{(n)}(a) = \lim_{x\to a}T_n^{(n)}(x) = \lim_{x\to a}f^{(n)}(x) = f^{(n)}(a)$$

We use the fact that both $T_n^{(n)}$ and $f^{(n)}$ are continuous at $x = a$.

Exercise 5.2.6

Finish the proof.

Exercise 5.2.7

Prove the converse.

Exercise 5.2.8

Prove that the nth degree Taylor polynomial of an nth degree polynomial is that polynomial.

Warning!

These are numbers, not functions; the derivatives are evaluated at $x = a$:

$$c_k = \frac{1}{k!}f^{(k)}\Big|_{x=a}\,.$$

So, we have:

$$T_0 = f(a),\quad T_{n+1}(x) = T_n + \frac{1}{(n+1)!}f^{(n+1)}(a)(x-a)^{n+1}\,,$$

and, in sigma notation:

Theorem 5.2.9: Taylor Polynomials of Functions

Suppose f is n times continuously differentiable at $x = a$. Then the n-th Taylor polynomial of f is the following:

$$\boxed{T_n(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(a)(x-a)^k}$$

This is indeed a polynomial of degree n. It is centered at a.

Example 5.2.10: exponent at 0

Some functions are so easy to differentiate that we can quickly find *all* of its Taylor polynomials. For example, consider

$$f(x) = e^x$$

at $x = 0$. Then

$$f^{(k)}(0) = e^x \Big|_{x=0} = 1\,.$$

Therefore,

$$T_n(x) = \sum_{k=0}^{n} \frac{1}{k!} x^k\,.$$

Then, we have an approximation formula:

$$e \approx 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + ... + \frac{1}{n!}\,,$$

which gives a better accuracy with each new term. The exact meaning of this statement is explained later.

The recursive formula for the Taylor polynomials is especially convenient:

$$T_{n+1}(x) = T_n(x) + \frac{1}{(n+1)!}(x-a)^{n+1}\,.$$

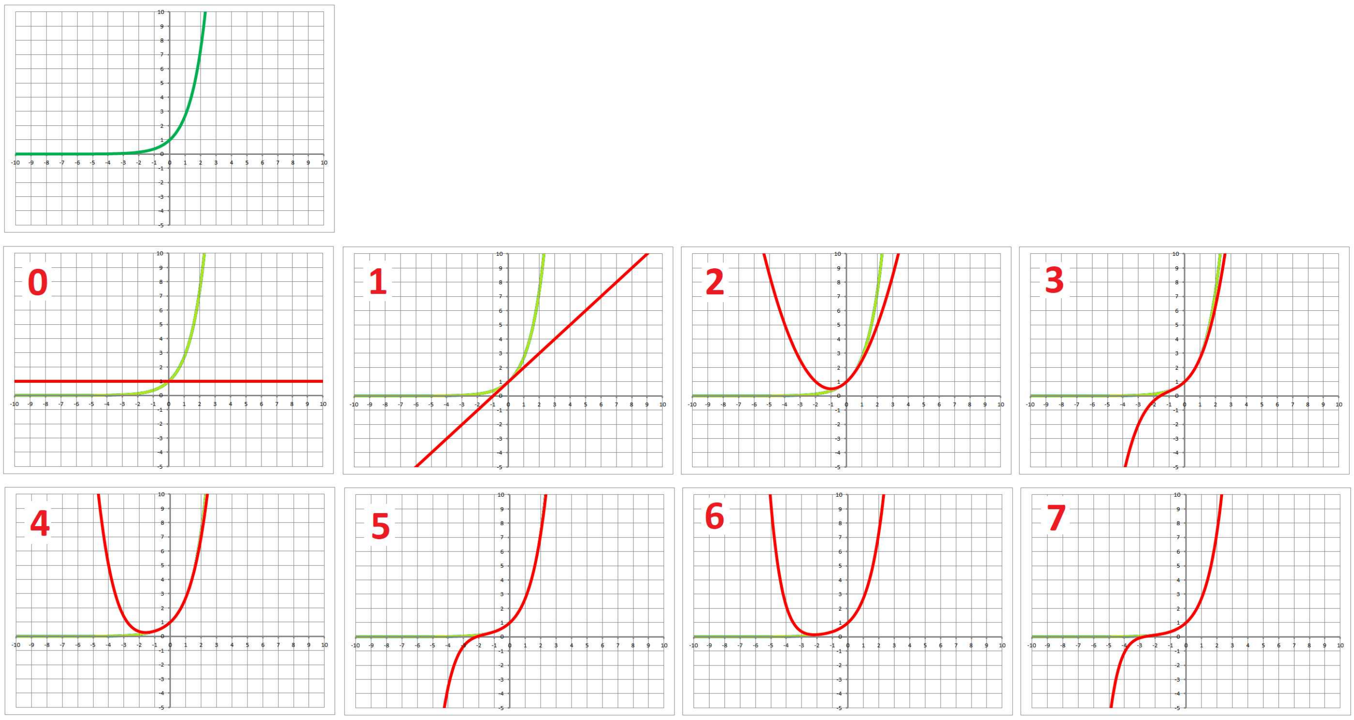
We use it in a spreadsheet as follows:

```
=RC[-1]+R8C/FACT(R7C)*(RC1-R4C2)^R7C
```

The first three approximations are:

$$T_0(x) = 1,\ T_1(x) = x + 1,\ T_2(x) = \frac{1}{2}x^2 + x + 1\,.$$

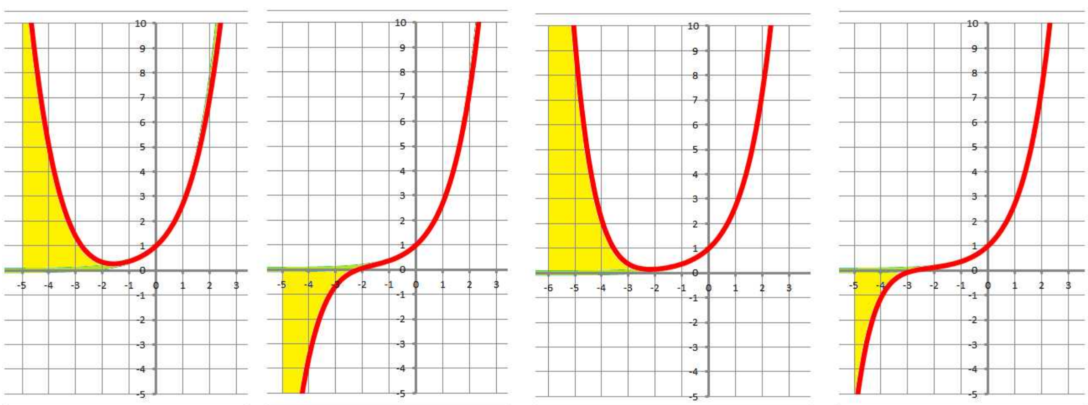
We can create as many as we like in each column:



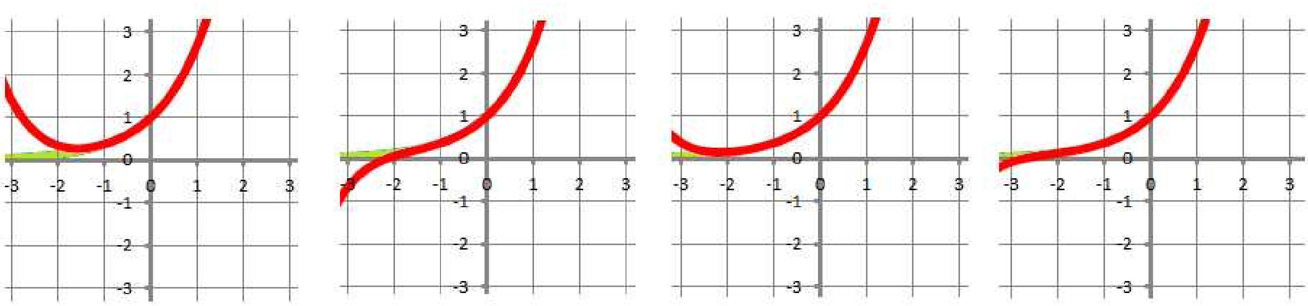
We can see how the curves start to resemble the original graph – but only in the vicinity of $x = 0$. Elsewhere, we know that polynomials have the property that

$$T_n(x) \to \infty \quad \text{as } x \to \infty \,.$$

Therefore, they can never get close to the horizontal asymptote for $x \to -\infty$ (degrees $4 - 7$):



We can't see it on the other end, but, for $x \to +\infty$, we know from L'Hopital's Rule that polynomials are too slow to compete with the exponential function. The good news is that, within an interval centered at 0, there is virtually no difference between the two graphs (degrees $4 - 7$):



We have, by far, more than just a sequence of approximations.

Recall from Volume 1 that as numbers are replaced with their approximations, we are still able to do algebra with them. Similarly, as functions are replaced with their approximations, we are still able to do *calculus* with them:

Theorem 5.2.11: Differentiation of Taylor Polynomials

Suppose f is n times continuously differentiable at $x = a$. Then the first n derivatives of the nth Taylor polynomial of a function f agree with those of the function f itself; i.e.,

$$\boxed{T_n^{(n)}(a) = f^{(n)}(a)}$$

Conversely, this polynomial is the only one that satisfies this property.

Proof.

Since T_n is a polynomial of degree n, the only term that matters for its nth derivative is the last one, $c_n(x-a)^n$. By the theorem above, we have:

$$T_n^{(n)}(x) = \left(c_n(x-a)^n\right)^{(n)} = c_n \cdot n! = \frac{f^{(n)}(a)}{n!} \cdot n! = f^{(n)}(a)\,.$$

Exercise 5.2.12

Prove the converse.

Exercise 5.2.13

Prove that the Taylor polynomials of an even (odd) function have only even (odd) terms.

Example 5.2.14: exponent at 1

Let's again consider

$$f(x) = e^x$$

but at $x = 1$ this time.

We still have all the derivatives ready:

$$f^{(k)}(1) = e^x \Big|_{x=1} = e\,.$$

Therefore, the Taylor polynomial is:

$$S_n(x) = \sum_{k=0}^{n} \frac{e}{k!}(x-1)^k\,.$$

It is named this way in order to compare it to the Taylor polynomial of the same function at $x = 0$:

$$T_n(x) = \sum_{k=0}^{n} \frac{1}{k!}x^k\,.$$

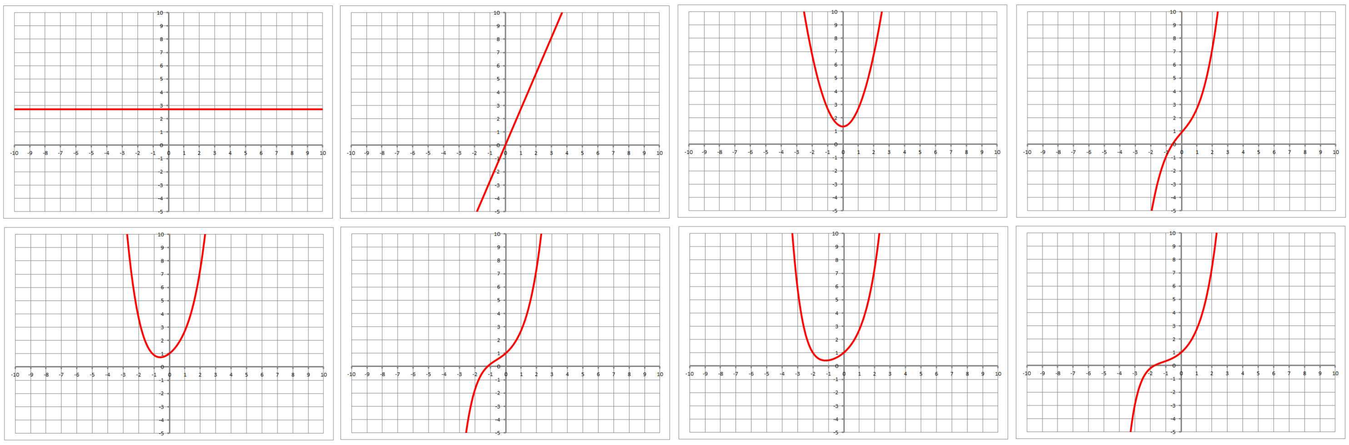
So, we have a simple relation between them:

$$S_n(x) = eT_n(x-1)\,.$$

The original is shifted one unit right and then stretched vertically by a factor of e... The first three approximations are:

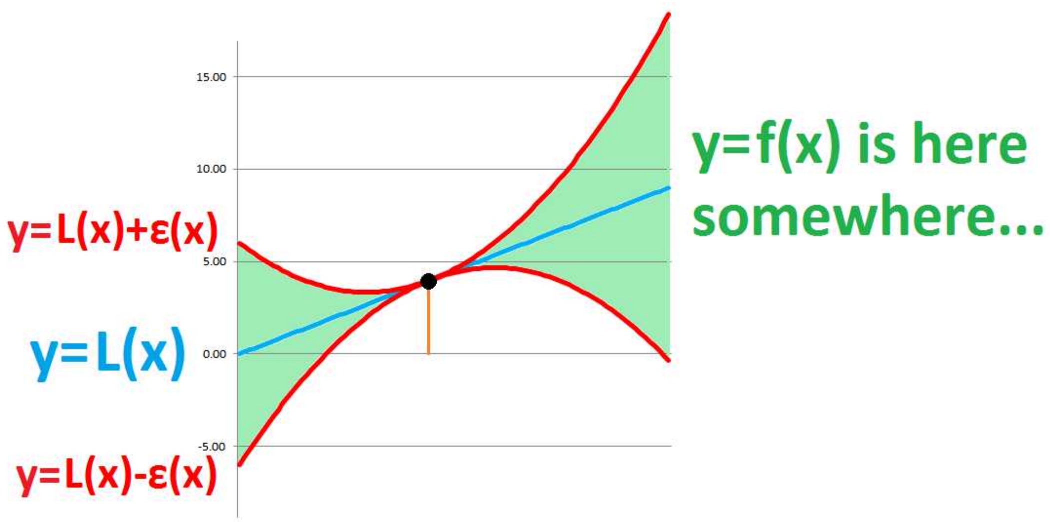
$$S_0(x) = e,\ S_1(x) = e(x-1) + e,\ S_2(x) = \frac{e}{2}(x-1)^2 + e(x-1) + e\,.$$

We see approximations of degrees $0 - 7$ below:



How well do our approximations work?

Recall how a linear approximation of a function f comes with a "funnel": it is centered on the tangent line, and its upper and lower bounds are the two parabolas (just like quadratic approximations). This is where the unknown function f must reside:



Similarly to the error bounds for Riemann integrals (Chapter 3), we have intervals that contain the unknown value – if we fix a value of x.

For an approximation of a higher degree, the blue line in the illustration is to be replaced with a parabola or a higher degree polynomial. We accept the result below without proof:

Theorem 5.2.15: Error Bound for Taylor Approximation

Suppose a function f is $(n+1)$ times differentiable at $x = a$. Suppose also that for each $i = 0, 1, 2, ..., n+1$, we have

$$|f^{(i)}(t)| < K_i \text{ for every } t \text{ between } a \text{ and } x\,,$$

and some real number K_i. Then

$$E_n(x) = |f(x) - T_n(x)| \leq K_{n+1} \frac{|x-a|^{n+1}}{(n+1)!}\,.$$

The most important consequence is the following:

Corollary 5.2.16: Taylor Approximation

Suppose a function f is infinitely many times differentiable at $x = a$. Suppose also that for each $i = 0, 1, 2, \ldots$, we have

$$|f^{(i)}(t)| < K_i \text{ for every } t \text{ between } a \text{ and } x\,,$$

and some real number K_i. Then the values of the Taylor polynomials at x converge to the value of the function at x:

$$\boxed{T_n(x) \to f(x) \quad \text{as } n \to \infty}$$

The result will serve as a cornerstone for the rest of the theory.

Exercise 5.2.17

Derive the corollary from the theorem.

Just as with the bounds for integrals, we need to know bounds for the derivatives of the unknown function f – a priori – in order to know how far it can go from the value that we do know. Note also that the nth error bound resembles the $(n+1)$st Taylor term.

The inequality is a "squeeze" (Chapters 2DC-1 and 2DC-2) for the unknown function f:

$$T_n(x) - E_n(x) < f(x) < T_n(x) + E_n(x)\,.$$

This inequality guarantees that:

- The error approaches to 0 as $n \to \infty$ for each x.
- The error approaches to 0 as $x \to a$ for each n.

In fact, the latter convergence is *fast*: quadratic, cubic, etc.

Example 5.2.18: root

Let's see how close we are to the truth with our quadratic approximation of $f(x) = \sqrt{x}$ around $a = 4$ from the last section. Recall that we have:

$$T_2(x) = 2 + \frac{1}{4}(x-4) - \frac{1}{2 \cdot 32}(x-4)^2\,.$$

The result comes from our computations of the derivatives up to the second to which we add the third:

$$\begin{array}{llllll}
f(x) & = \sqrt{x} & & \Longrightarrow & f(4) & = 2 \\
f'(x) & = (x^{1/2})' & = \dfrac{1}{2}x^{-1/2} & \Longrightarrow & f'(4) & = \dfrac{1}{4} \\
f''(x) & = \left(\dfrac{1}{2}x^{-1/2}\right)' & = -\dfrac{1}{4}x^{-3/2} & \Longrightarrow & f''(4) & = -\dfrac{1}{32} \\
f^{(3)}(x) & = \left(-\dfrac{1}{4}x^{-3/2}\right)' & = -\dfrac{1}{4}\dfrac{-3}{2}x^{-5/2} & = \dfrac{3}{8}x^{-5/2} & &
\end{array}$$

Next, we notice that $f^{(3)}$ is decreasing. Therefore, over the interval $[4, +\infty)$, our best (smallest) upper bound for it is its initial value:

$$|f^{(3)}(x)| \le |f^{(3)}(4)| = \frac{3}{8}4^{-5/2} = \frac{3}{8 \cdot 32} = \frac{3}{256}\,.$$

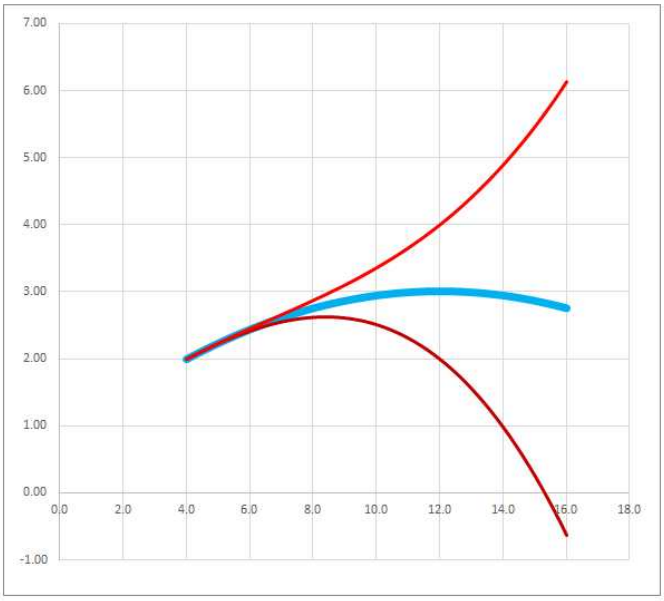
So, our best (smallest) choice of the upper bound is:

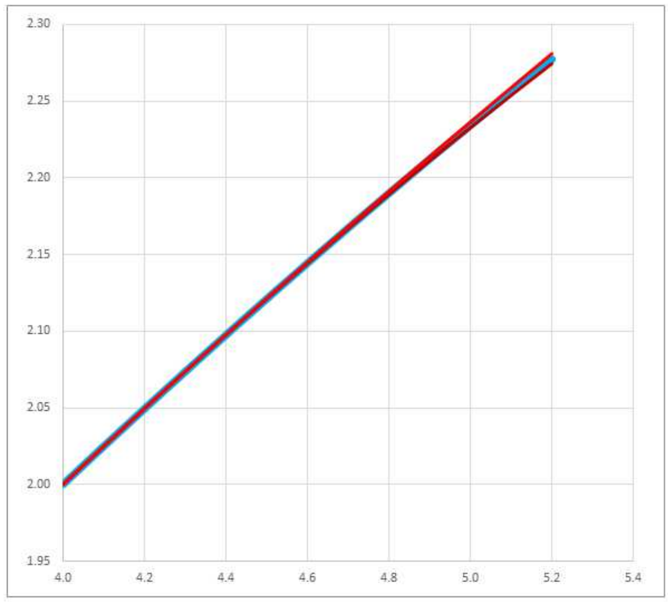
$$K_3 = \frac{3}{256}\,.$$

Then,

$$E_2(x) = |f(x) - T_2(x)| \le K_3 \frac{|x-4|^3}{3!} = \frac{3}{256 \cdot 3!}|x-4|^3 = \frac{1}{512}|x-4|^3 .$$

This is where the graph of $y = \sqrt{x}$ lies:





We can now address the accuracy of our approximation of $\sqrt{4.1}$:

$$E_2(4.1) = \frac{1}{512}|4.1-4|^3 = \frac{0.001}{512} \approx 0.000002 .$$

Therefore, we have an interval where the true value of $\sqrt{4.1}$ must lie:

$$T_2(4) - .000002 \le \sqrt{4.1} \le T_2(4) + .000002 .$$

Example 5.2.19: exponent

Let's estimate $e^{-0.01}$ within 6 decimals. In other words, we need to find such an n that we are guaranteed to have:

$$\left| e^{-0.01} - T_n(-.01) \right| < 10^{-6} ,$$

where T_n is the nth Taylor polynomial of e^x around $x = 0$. We estimate the derivative of this function on the interval $[-0.01, 0]$:

$$\left| (e^x)^{(i)} \right| = e^x \le 1 = K_i .$$

Then, how do we make the error bound satisfy this:

$$\left| e^{-0.01} - T_n(-.01) \right| \le K_{n+1} \frac{|x-a|^{n+1}}{(n+1)!} < 10^{-6} ?$$

We re-write and solve this inequality for n:

$$\frac{0.1^{n+1}}{(n+1)!} < 10^{-6} .$$

A large enough n will work. We simply go through a few values of $n = 1, 2, 3, ...$ until the inequality is satisfied:

n	3	4	5
$\frac{0.1^{n+1}}{(n+1)!}$	0.000004167	0.000000083	0.000000001

It's $n = 4$ and, therefore, the answer is:

$$T_4(-0.01) = \sum_{i=0}^{4} \frac{1}{i!}(-0.01)^i .$$

We now come back to the requirement from last section that the approximations have to provide faster and faster convergence to zero of the error.

Corollary 5.2.20: Error Convergence

Suppose a function f is $(n+1)$ times differentiable at $x=a$. Suppose also that for each $i=0,1,2,...,n+1$, we have

$$|f^{(i)}(t)| < K_i \text{ for every } t \text{ between } a \text{ and } x\,,$$

and some real number K_i. Then

$$\frac{E_n(x)}{|x-a|^n} = \frac{|f(x)-T_n(x)|}{|x-a|^n} \to 0 \text{ as } x \to a\,.$$

5.3. Sequences of functions

Let's take a closer look at the theorems in the last section.

- First, the Taylor polynomials of f form a *sequence of functions*

$$T_n:\ n=0,1,2,3,...$$

- Second, suppose we fix a value x within the interval, then we have a *sequence of numbers*

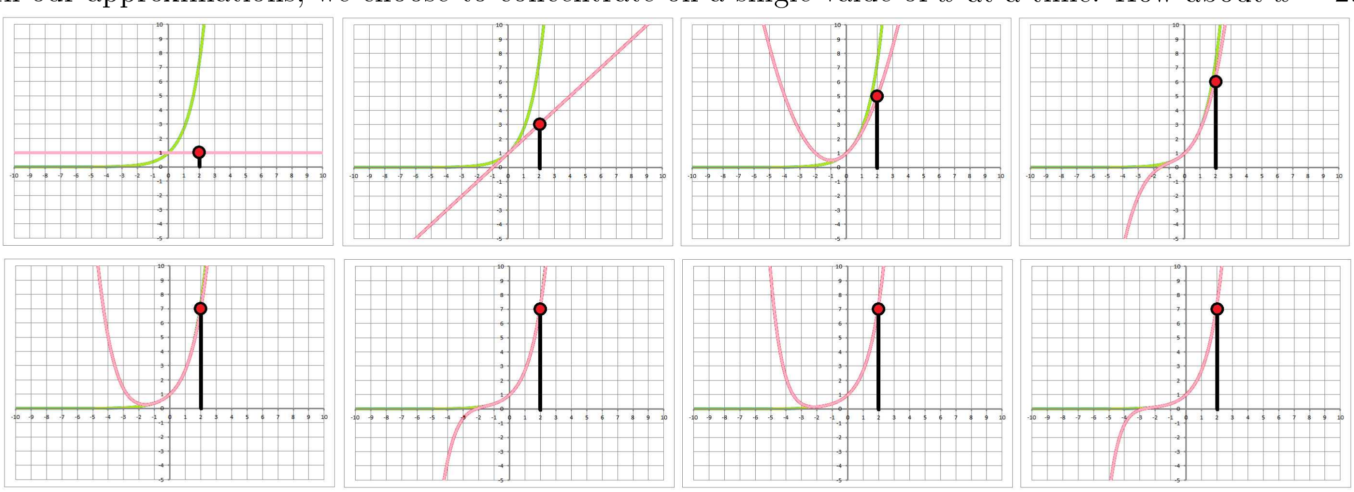
$$T_n(x):\ n=0,1,2,3,...$$

This sequence has the following property according to the theorem about *Taylor Approximation* in the last section, for each x:

$$\boxed{T_n(x) \to f(x) \ \text{ as } n \to \infty}$$

Example 5.3.1: exponent

In our approximations, we choose to concentrate on a single value of x at a time. How about $x=2$:



From this sequence of approximations of e^x, we take this:

$$T_n(2) \to e^2 \ \text{ as } \ n \to \infty\,.$$

And so on.

Since this convergence occurs for *each* x, we can speak of convergence of the whole sequence of Taylor polynomials. The general idea is the following.

Definition 5.3.2: pointwise convergent sequence of functions

Suppose we have a sequence of functions f_n defined on interval I. We say that the sequence *converges pointwise on* I to a function f if for every x, the values of these functions at x converge, as a sequence of numbers, to the value of f at x, i.e.,

$$f_n(x) \to f(x)\,.$$

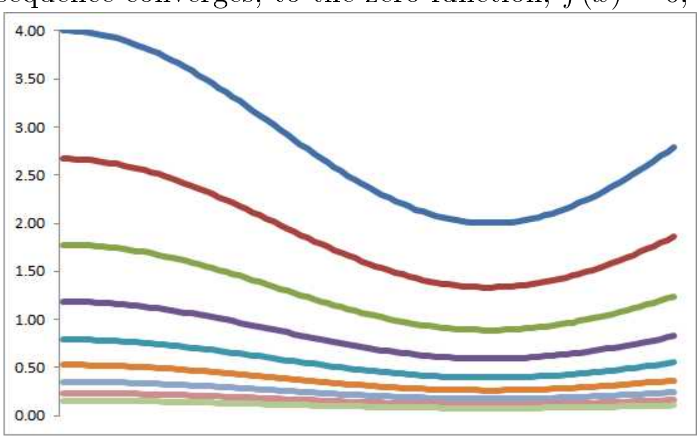
Otherwise, we say that the sequence *diverges pointwise*.

Warning!

It only takes divergence for a single value of x to make the whole sequence of functions diverge.

Example 5.3.3: shrinking

This is how, typically, a sequence converges, to the zero function, $f(x) = 0$, in this case:



The functions are constructed as follows:

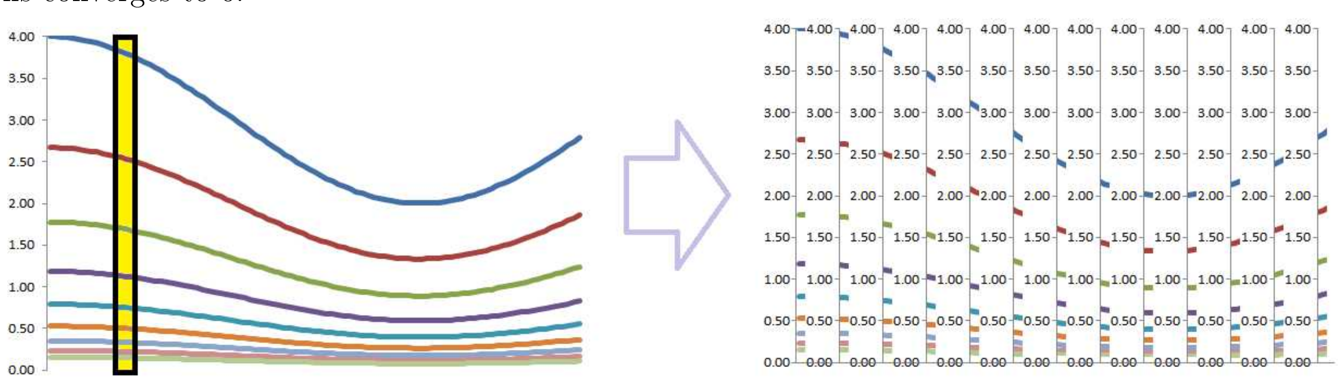
$$g(x) = 3 - \cos x, \quad f_n(x) = \left(\frac{2}{3}\right)^n g(x)\,.$$

You can choose any other function g.

The multiple is getting smaller and shrinks the graph of the function g toward the x-axis. The proof of convergence is routine (after all, the sequence is geometric):

$$|f_n(x)| = \left|\left(\frac{2}{3}\right)^n g(x)\right| = |g(x)|\left|\left(\frac{2}{3}\right)^n\right| \to 0\,,$$

by the *Constant Multiple Rule*. Thus, each numerical sequence produced from this sequence of functions converges to 0:



They do this independently of each other.

Example 5.3.4: shifting

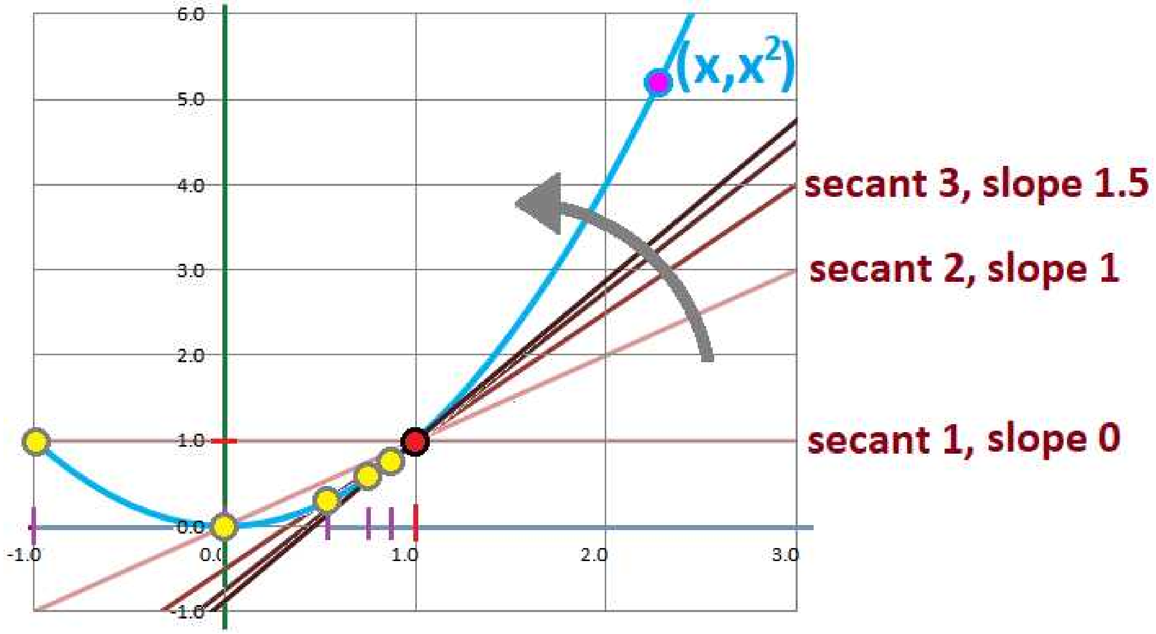
Another simple choice of a sequence is:

$$f_n(x) = f(x) + \frac{1}{n} \to f(x) + 0 = f(x)\,,$$

for each x, no matter that f is.

Example 5.3.5: secants and tangents

We said, informally, that the secant lines converge to the tangent line (Chapter 2DC-3):



Let's interpret the convergence of the secant lines to the tangent line as convergence of functions. Suppose a function f is defined on an open interval that contains $x = a$. For each $n = 1, 2, 3, ...$, define a function f_n as the linear function that passes through these two points:

$$(a, f(a)) \text{ and } \left(a + \frac{1}{n}, f\left(a + \frac{1}{n}\right)\right).$$

Then,

$$f_n(x) = f(a) + \frac{f(a + 1/n) - f(a)}{1/n}(x - a)\,.$$

If f is differentiable at $x = a$, the fraction, its slope, converges to $f'(a)$ as $n \to \infty$ according to the definition of the derivative:

$$\frac{f(a + 1/n) - f(a)}{1/n} \to f'(a)\,.$$

Therefore, we have:

$$\begin{array}{ccccccc} f_n(x) & = & f(a) & + & \dfrac{f(a+1/n) - f(a)}{1/n} & & (x-a) \\ \downarrow & & || & & \downarrow & & || \\ f(x) & = & f(a) & + & f'(a) & & (x-a) \end{array}$$

This new function is the best linear approximation of the f at a and its graph is the tangent line of f at a. When the function f is not differentiable at this point, our sequence of functions diverges.

Example 5.3.6: Riemann sums and Riemann integrals

We said that the Riemann sums (say left end) converge to the Riemann integral (Chapter 1):

$$L_n \to \int_a^b f\,dx\,.$$

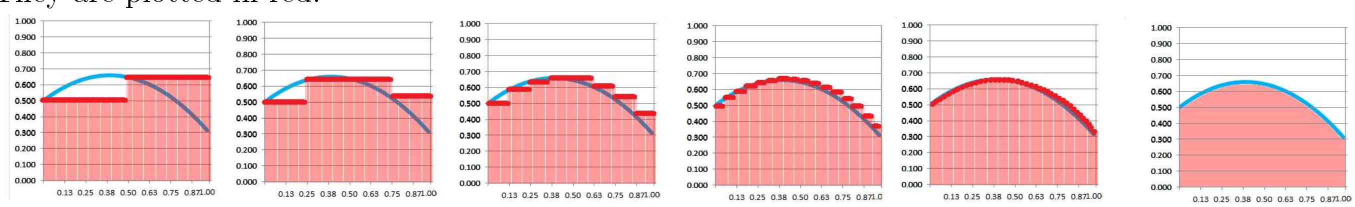
There is nothing imprecise about saying that.

However, that was a sequence of *numbers*. Is there a sequence of *functions* here? Yes, the sequence of step-functions that represent the left-end Riemann sums.

Let's make this specific. Suppose f is defined on interval $[a,b]$. Suppose also that the interval is equipped with a specific augmented partition, the left-end, for each $n = 0,1,2,3,...$, with $\Delta x = (b-a)/n$ and $x_i = a + \Delta x \cdot i$. Then, for each $n = 0,1,2,3,...$, define a step-function piecewise as follows:

$$f_n(x) = F(x_i) \text{ when } x_i \le x < x_{i+1}\,.$$

They are plotted in red:



When f is continuous, we know that this sequence will converge to f pointwise.

Exercise 5.3.7

Prove the last statement.

Exercise 5.3.8

(a) Show that when the function is not continuous on this interval, our sequence of functions might diverge. (b) Show that it doesn't *have to* diverge, however.

Exercise 5.3.9

Consider also the sequences of step-functions that represent the right and middle point Riemann sums and the trapezoid approximations.

In all these examples, the convergence of functions means convergence of (many) sequences of *numbers*. However, the whole graphs of these functions f_n seem to completely accumulate toward the graph of f. This is a "stronger" kind of convergence:

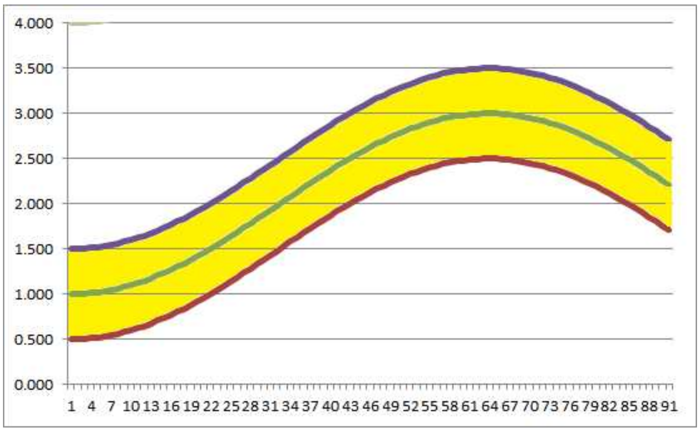
Definition 5.3.10: converges uniformly

Suppose we have a sequence of functions f_n defined on interval I. We say that the sequence *converges uniformly on* I to function f if the maximum value of the difference converges, as a sequence of numbers, to zero:

$$\max_I |f_n(x) - f(x)| \to 0\,.$$

Otherwise, we say that the sequence *diverges uniformly*.

In other words, the graphs of the functions of the sequence will *eventually* fit entirely within a strip around f, no matter how narrow:



It is clear that every uniformly convergent sequence converges pointwise (to the same function):

$$\sup_I |f_n(x) - f(x)| \to 0 \implies |f_n(x) - f(x)| \to 0 \text{ for each } x \text{ in } I\,.$$

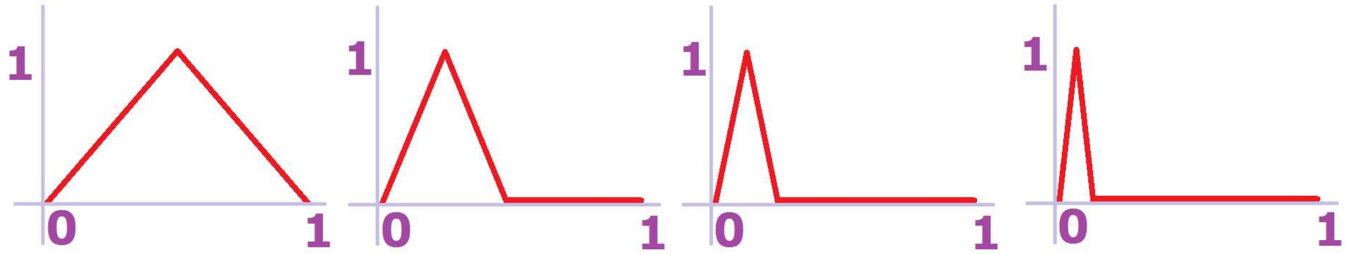
Exercise 5.3.11

Prove the last statement.

The converse isn't true.

Example 5.3.12: tooth sequence

We have a sequence of continuous functions defined on $I = [0,1]$ that satisfies: $f_n(0) = 0$ and $f_n(x) = 0$ for $x > \frac{1}{n+1}$ and the rest of the values produce a "tooth" of height 1:

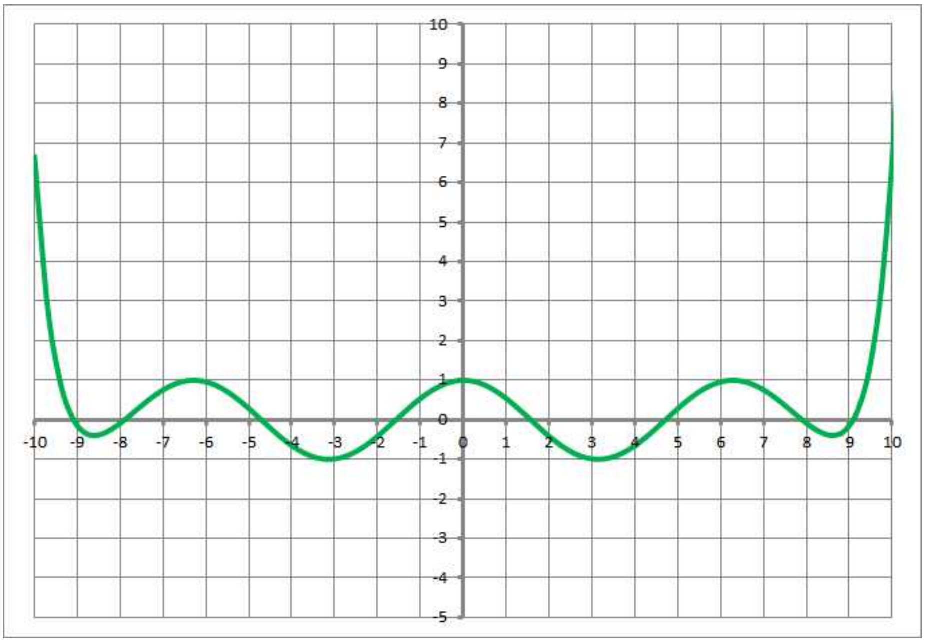


Then, we have:

- f_n converges to 0, the function, pointwise because $f_n(x)$ eventually becomes zero for each x.
- f_n does not converge to 0, the function, uniformly because the tooth will always stick out of any strip narrower than 1.

Example 5.3.13: Taylor polynomials

From the last section, we conclude that, whenever the derivatives of f are known to be bounded on a bounded interval $I = [a,b]$, the sequence of Taylor polynomials converges to f uniformly on I. On an unbounded interval, the uniform convergence isn't guaranteed; for example, consider $f(x) = \cos x$ on $(-\infty, +\infty)$:



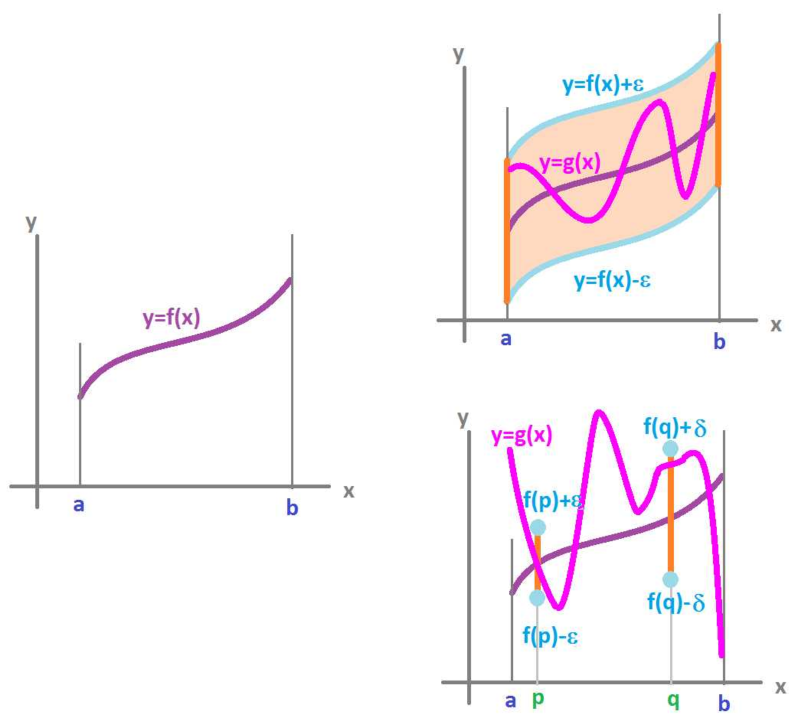
The reason is that a polynomial of any degree above 0 will eventually go to infinity (Volume 2). Pointwise convergence remains.

Let's compare the two types of convergence again by referring to the original definition of convergence:

- Sequence f_n converges to f pointwise if
  - ▶ *for each* x, for any $\varepsilon > 0$ there is such an $N > 0$ that
$$n > N \implies |f_n(x) - f(x)| < \varepsilon\,.$$
- Sequence f_n converges to f uniformly if
  - ▶ for any $\varepsilon > 0$ there is such an $N > 0$ that, *for each* x,
$$n > N \implies |f_n(x) - f(x)| < \varepsilon\,.$$

As you see, we just moved "for each x" within the sentence.

Here is how different the idea of "close" is relative to these two types of convergence:



Exercise 5.3.14

Investigate the convergence of the sequence $f_n(x) = \dfrac{1}{nx}$.

5.4. Infinite series

Warning!

Series are sequences.

Our goal is to be able to find out the extent to which our approximations – polynomial approximations of functions – work.

Specifically, we need to know *for what values of* x the sequence of the values of the Taylor polynomials $T_n(x)$ of a function f converges to the value of the function $f(x)$. We will also estimate the error, just as we did earlier in this chapter.

Following the development in the last section, we can consider sequences of *any* functions as approximations. What makes Taylor polynomials different? It's in the *recursive* formula for the Taylor polynomials, for a fixed x:

$$T_{n+1}(x) = T_n(x) + c_{n+1}(x-a)^{n+1}\,.$$

The formula continues to compute the Taylor polynomials of higher and higher degrees by simply adding new terms to the previous result. Taylor polynomials are just a special case of the following:

Definition 5.4.1: power series centered at point

A sequence q_n of polynomials given by a recursive formula:

$$q_{n+1}(x) = q_n(x) + c_{n+1}(x-a)^{n+1},\ n = 0, 1, 2, \ldots,$$

for some fixed number a and a sequence of coefficients c_n, is called a *power series centered at* a.

Power series may come fully formed.

How do we add together the infinitely many terms of such a sequence? *Via limits* is the only answer.

The most important definition of this chapter repeats the one in Chapter 1:

Definition 5.4.2: sequence of sums, partial sums

Suppose

$$a_n:\ n = s, s+1, s+2, \ldots$$

is a sequence. Its *sequence of sums*

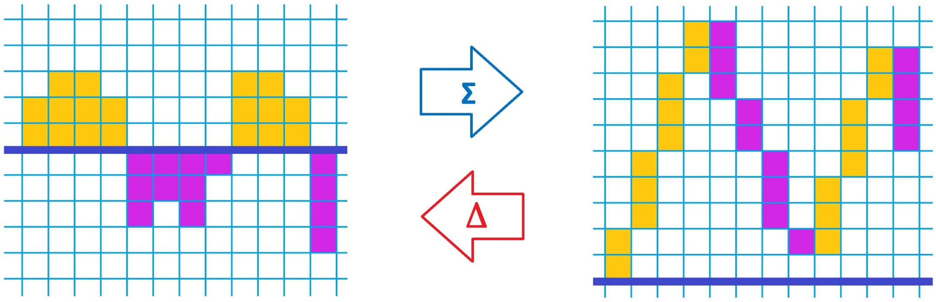
$$p_n:\ n = s, s+1, s+2, \ldots$$

is the sequence defined by the following recursive formula:

$$p_s = a_s, \quad p_{n+1} = p_n + a_n\,.$$

The sequence is also called the *partial sums* of the original.

This process is a familiar way of creating new sequences from old (Volume 2). Imagine that we stack the elements of the original sequence a_n (left) on top of each other to produce p_n, a new sequence (right):



In the next several sections, we will concentrate on such sequences of *numbers* (rather than functions) and occasionally apply the results to power series.

The original sequence is left behind, and it is *the limit of this new sequence* that we are after.

Example 5.4.3: limit of sums

In either of the two tables below, we have a sequence given in the first two columns. Its nth term formula is known and, because of that, its limit is also easy to find. The third column shows the sequence of (partial) sums of the first. Its nth term formula is unknown and, because of that, its limit is not easy to find.

n	a_n	p_n
1	$\frac{1}{1}$	$\frac{1}{1}$
2	$\frac{1}{2}$	$\frac{1}{1}+\frac{1}{2}$
3	$\frac{1}{3}$	$\frac{1}{1}+\frac{1}{2}+\frac{1}{3}$
$\vdots$	$\vdots$	$\vdots$
n	$\frac{1}{n}$	$\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+...+\frac{1}{n}$, formula?
$\vdots$	$\vdots$	$\vdots$
$\downarrow$	$\downarrow$	$\downarrow$
∞	0	?

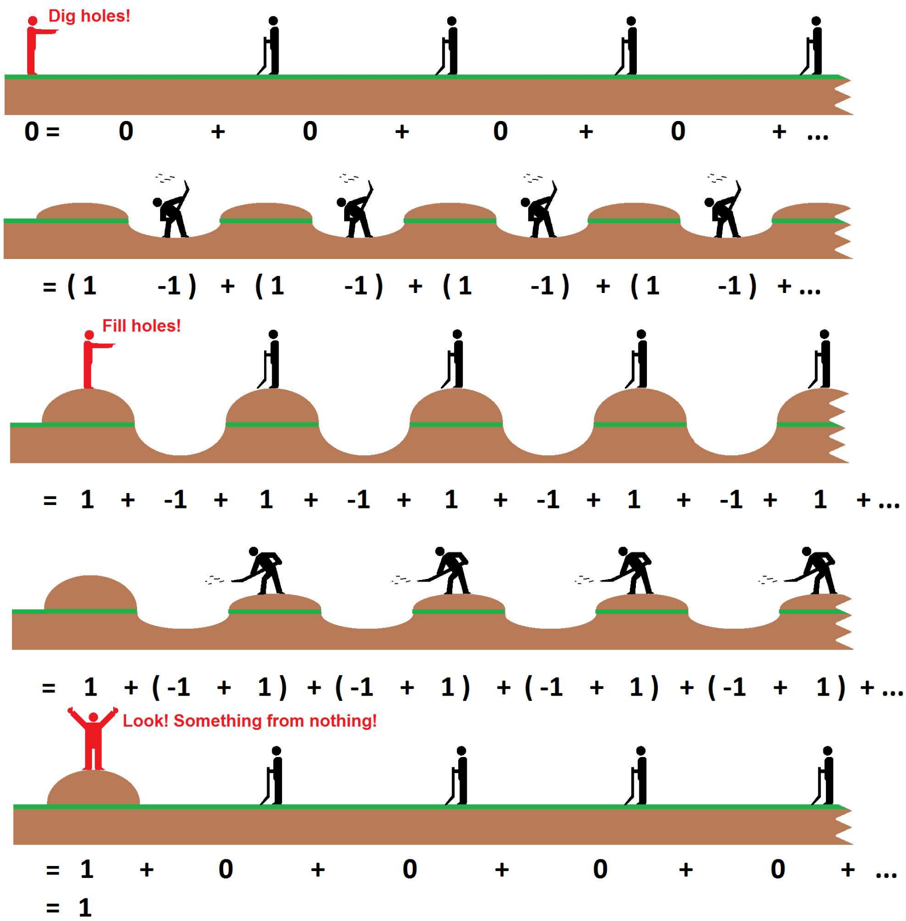
n	a_n	p_n
1	$\frac{1}{1}$	$\frac{1}{1}$
2	$\frac{1}{2}$	$\frac{1}{1}+\frac{1}{2}$
3	$\frac{1}{4}$	$\frac{1}{1}+\frac{1}{2}+\frac{1}{4}$
$\vdots$	$\vdots$	$\vdots$
n	$\frac{1}{2^n}$	$\frac{1}{1}+\frac{1}{2}+\frac{1}{4}+...+\frac{1}{2^{n-1}}$, formula?
$\vdots$	$\vdots$	$\vdots$
$\downarrow$	$\downarrow$	$\downarrow$
∞	0	?

Example 5.4.4: something from nothing

This example from Volume 2 (Chapter 2DC-1) shows what can happen if we ignore the issue of convergence:

$$\begin{array}{lllllllllll}
0 & \stackrel{?}{=} & 0 & & +0 & & +0 & & +0 & & +... \\
 & \stackrel{?}{=} & (1 & -1) & +(1 & -1) & +(1 & -1) & +(1 & -1) & +... \\
 & \stackrel{?}{=} & 1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 & +... \\
 & \stackrel{?}{=} & 1 & +(-1 & +1) & +(-1 & +1) & +(-1 & +1) & +(-1 & +1) & +... \\
 & \stackrel{?}{=} & 1 & +0 & & +0 & & +0 & & +0 & +... \\
 & \stackrel{?}{=} & 1. & & & & & & & &
\end{array}$$

Removing the parentheses in an infinite computation is not what we would do anymore – in light of our definition. The result qualifies as an example of “something from nothing”:



In the "computation", we go – implicitly – through *three* sequences. The fact that the first and the third both converge obscures the fact that the second diverges. To detect the switches, we observe that the sequences that produce the three sums are different! We list the first two below:

n	a_n	p_n	$=$	p_n
1	$0=1-1$	0	$=$	0
2	$0=1-1$	$0+0$	$=$	0
3	$0=1-1$	$0+0+0$	$=$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	$0=1-1$	$0+0+0+...+0$	$=$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\downarrow$	$\downarrow$			$\downarrow$
∞	0			0

n	a_n	p_n	$=$	p_n
1	1	1	$=$	1
2	-1	$1-1$	$=$	0
3	1	$1-1+1$	$=$	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	$(-1)^n$	$1-1+1-...+(-1)^n$	$=$	$1,\ 0$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\downarrow$	$\downarrow$			$\downarrow$
∞	DNE			DNE

As a divergent sequence, the second one cannot be equal to *anything*.

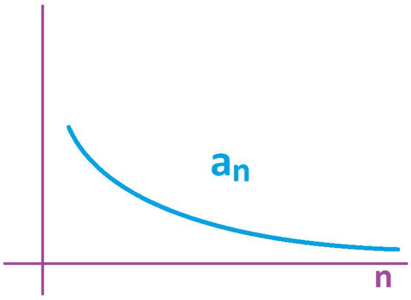
Exercise 5.4.5

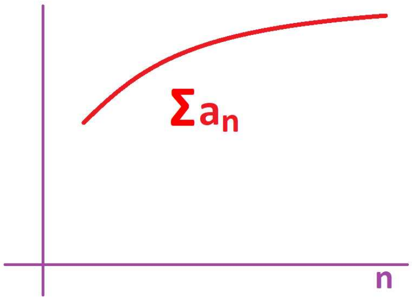
Which of the "$\overset{?}{=}$" signs above is incorrect?

In summary,

▶ *A series is a pair of sequences.*

The first produces the second via summation, for example:





Is it even possible to find the limit of the sequence of sums p_n without first finding the formulas for its nth term? Formulas that have "..." don't count:

	list of terms	formula for nth term
original sequence:	$1,\ \frac{1}{2},\ \frac{1}{3},\ \frac{1}{4}, ...$	$\frac{1}{n}$
(partial) sums sequence:	$1,$ $1+\frac{1}{2},$ $1+\frac{1}{2}+\frac{1}{3},$ ... $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+...+\frac{1}{n}$	 $\sum_{k=1}^{n}\frac{1}{k}$

Neither do formulas that rely on the sigma notation.

The challenge posed by series is that they never come with an nth-term formula! As an example, an nth partial sum of a power series is a polynomial of x and, from what we know about polynomials (Chapter 1PC-4), it can only be simplified in a few very special cases, such as the Binomial Formula (Chapter 1PC-1).

Definition 5.4.6: sum of sequence or series

For a sequence a_n, the limit S of its sequence of (partial) sums p_n is called by the *sum of the sequence* or, more commonly, the *sum of the series*:

$$\lim_{n\to\infty}\sum_{i=s}^{n} a_i = S\,.$$

When the limit of partial sums exists, we say that the *series converges.* When the limit does not exist, we say that the *series diverges.* A special case of divergence is the divergence to (plus or minus) infinity and we say that the sum is *infinite*:

$$\lim_{n\to\infty}\sum_{i=s}^{n} a_i = \infty\,.$$

Once again, a series is a sequence built from another via recursive addition.

Warning!

The starting point, s, of summation doesn't affect convergence but does affect the sum when the series converges.

Example 5.4.7: convergence of power series

Note that a power series may converge for some values of x and diverge for others. For example, the series

$$1 + x + x^2 + \dots$$

- converges for $x = 0$ but
- diverges for $x = 1$.

We also demonstrate below that it converges for $x = 1/2$.

This is an abbreviated notation to write the limit of partial sums:

Sum of series

$$\sum_{i=s}^{\infty} a_i = \lim_{n\to\infty} \sum_{i=s}^{n} a_i$$

Warning!

The Riemann sum of a function isn't a series.

Recall that here Σ stands for "S" meaning "sum". This is how the notation is deconstructed:

Sigma notation for series

$$\begin{array}{ccccc}
\text{beginning} & \text{and end values for } k & & & \\
\downarrow & & & & \\
\infty & & & & \\
 & \sum & \left(\dfrac{1}{2^k} + \dfrac{1}{3^k} \right) & = & \dfrac{7}{2} \\
k=0 & & & & \\
 & & \uparrow & & \uparrow \\
 & & \text{a specific sequence} & & \text{a specific number, } \pm\infty, \text{ or "DNE"}
\end{array}$$

Warning!

In some sources, the word "series" might refer to one, or a combination of all, of these three:

- the original sequence a_n,
- its sequence of sums p_n, and
- the limit of the latter.

Example 5.4.8: fineteness

When there are only finitely many non-zero elements in the sequence, its sum is simple:

$$a_i = 0 \text{ for each } i > N \implies \sum_{i=1}^{\infty} a_i = \sum_{i=1}^{N} a_i\,.$$

The series converges to this number.

Example 5.4.9: divergent

Non-zero constant is the simplest example of a convergent sequence but a divergent series:

$$a_i = 1 \text{ for each } i \implies \sum_{i=1}^{\infty} a_i = \lim_{n\to\infty} \sum_{i=1}^{n} 1 = \lim_{n\to\infty} n = \infty\,.$$

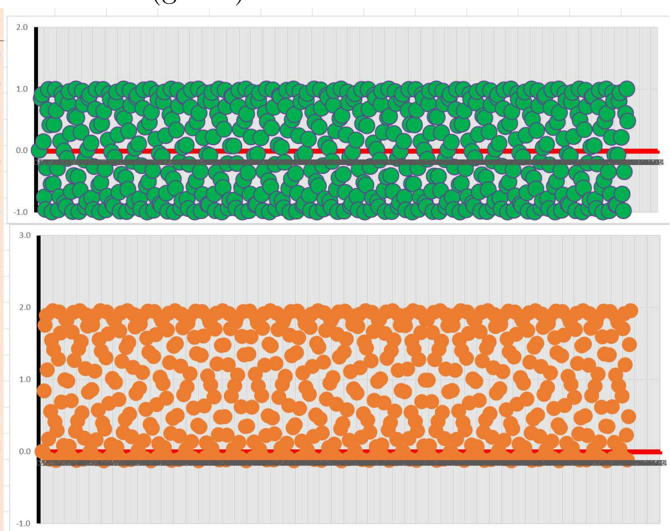
Example 5.4.10: $\sin n$ and random sequence

Consider the sequence

$$a_n = \sin n\,.$$

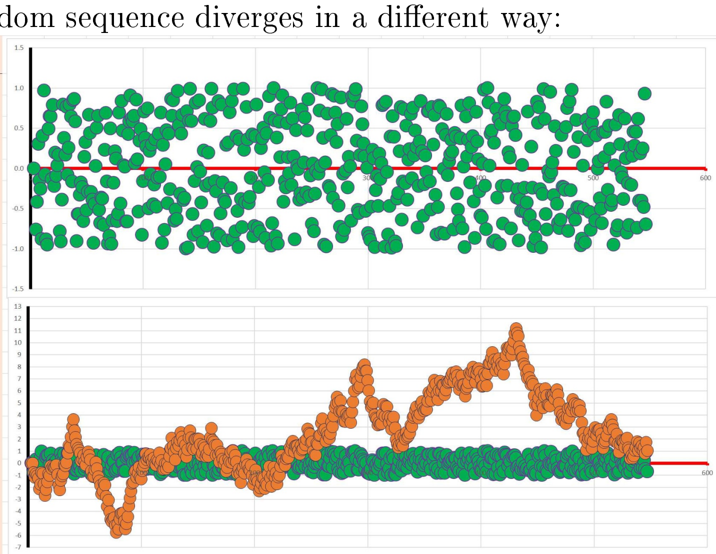
It looks uniformly spread between -1 and 1 (green):

n	a_n	sum
0		0.00
1	0.84	0.84
2	0.91	1.75
3	0.14	1.89
4	-0.76	1.14
5	-0.96	0.18
6	-0.28	-0.10
7	0.66	0.55
8	0.99	1.54
9	0.41	1.96
10	-0.54	1.41
11	-1.00	0.41
12	-0.54	-0.13
13	0.42	0.29
14	0.99	1.29
15	0.65	1.94
16	-0.29	1.65
17	-0.96	0.69
18	-0.75	-0.06



Its sequence of sums also looks uniformly spread between, 0 and 2 (orange). It, then, appears to diverge. A series of a truly random sequence diverges in a different way:

n	a_n	sum
0		0.00
1	-0.76	-0.76
2	-0.42	-1.18
3	0.30	-0.87
4	-0.15	-1.02
5	-0.25	-1.27
6	-0.88	-2.15
7	0.41	-1.75
8	0.97	-0.78
9	-0.08	-0.85
10	-0.88	-1.74
11	-0.95	-2.69
12	0.49	-2.19
13	0.65	-1.54
14	0.65	-0.90
15	-0.13	-1.03
16	0.79	-0.24
17	-0.22	-0.46
18	0.20	-0.26
19	-0.09	-0.36
20	0.04	-0.32
21	0.17	-0.15
22	-0.78	-0.93



We saw that, when facing infinity, algebra may fail. But it won't if the series converge! In that case, the series can be subjected to algebraic operations. Furthermore, the power series too can be subjected to calculus operations...

Example 5.4.11: decimal approximations

For a given real number, we can construct a series that tends to that number – via truncations of its decimal approximations. The sequence

$$a_n = 0.9,\ 0.09,\ 0.009,\ 0.0009,\ \ldots \text{ tends to } 0\,.$$

But its sequence of partial sums

$$p_n = 0.9,\ 0.99,\ 0.999,\ 0.9999, \ldots \text{ tends to } 1\,.$$

The sequence

$$a_n = 0.3,\ 0.03,\ 0.003,\ 0.0003,\ \ldots \text{ tends to } 0\,.$$

But its sequence of partial sums

$$p_n = 0.3,\ 0.33,\ 0.333,\ 0.3333,\ \ldots \text{ tends to } 1/3\,.$$

The idea of series then helps us understand infinite decimals.

- What is the meaning of 0.9999...? It is the sum of the series:

$$\sum_{i=1}^{\infty} 9 \cdot 10^{-i}\,.$$

- What is the meaning of 0.3333...? It is the sum of the series:

$$\sum_{i=1}^{\infty} 3 \cdot 10^{-i}\,.$$

Exercise 5.4.12

Find such a series for $1/6$.

We know that a convergent sequence can have only one limit.

Theorem 5.4.13: Uniqueness of Sum of Series

A series can have only one sum (finite or infinite).

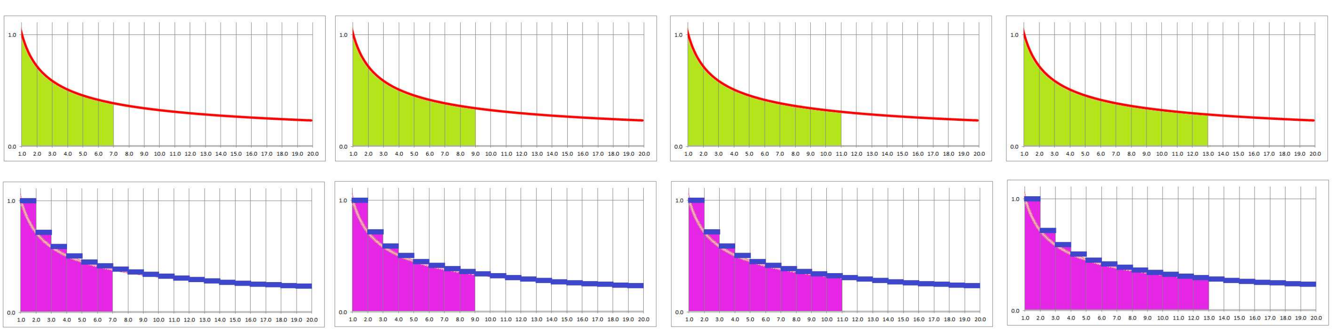
Thus, we are justified to speak of *the* sum.

This conclusion makes possible the theory we shall develop:

▶ Any power series defines a function with its domain consisting of those values of x for which the series converges.

Example 5.4.14: series vs. improper integrals

The way we use the limits to transition from sequences to series is familiar. It is identical to the transition from a function f to its integral over $[0, \infty)$:



Indeed, compare:

$$\begin{aligned} \int_1^{\infty} f(x)\,dx &= \lim_{b\to\infty} \int_1^b f(x)\,dx \\ \sum_{i=1}^{\infty} a_i &= \lim_{n\to\infty} \sum_{i=1}^{n} a_i \end{aligned}$$

Furthermore, the latter will fall under the scope of the former if we choose f to be the step-function

produced by the sequence a_n:

$$f(x) = a_{[x]}\,.$$

Warning!

The worst mistake one can make is to confuse the limit of a_n with the limit of p_n.

5.5. Main classes of series

The key to evaluating sums of a series, or discovering that it diverges, is to find an *explicit formula for the nth partial sum* p_n of a_n.

We have to be careful and not take the former over the latter:

$$\boxed{\lim_{n\to\infty} a_n \quad \text{vs.} \quad \lim_{n\to\infty} p_n}$$

Example 5.5.1: constant

Recall about the *constant sequence* that, for any real c, we have

$$\lim_{n\to\infty} c = c\,.$$

The result tells us nothing about the series produced by the sequence! Consider instead:

$$c + c + c + \ldots = \lim_{n\to\infty} \sum_{k=1}^{n} c = \lim_{n\to\infty} nc\,.$$

Therefore, such a constant series diverges unless $c = 0$. The result is written as follows:

$$\sum_{k=1}^{\infty} c = \infty\,.$$

Example 5.5.2: arithmetic

More generally, consider what we know about *arithmetic progressions*; for any real numbers $m, b > 0$, we have:

$$\lim_{n\to\infty} (b + nm) = \begin{cases} -\infty & \text{if } m < 0\,, \\ b & \text{if } m = 0\,, \\ +\infty & \text{if } m > 0\,. \end{cases}$$

The result tells us nothing about the series! Instead, let's examine the partial sums and, because each is comprised of only finitely many terms, we can manipulate them algebraically before finding the limit, as follows:

$$b + (b+m) + (b+2m) + (b+3m) + \ldots \quad = \lim_{n\to\infty} \sum_{k=0}^{n} (b + km)\,.$$

Exercise 5.5.3

Show that such a series diverges unless $b = m = 0$:

$$\sum_{k=0}^{\infty}(b+km)=\infty\,.$$

There are more interesting series.

Definition 5.5.4: harmonic series

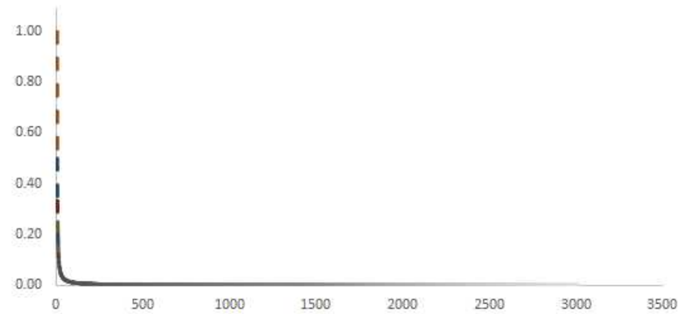
The series produced by the sequence of reciprocals $a_n = 1/n$,

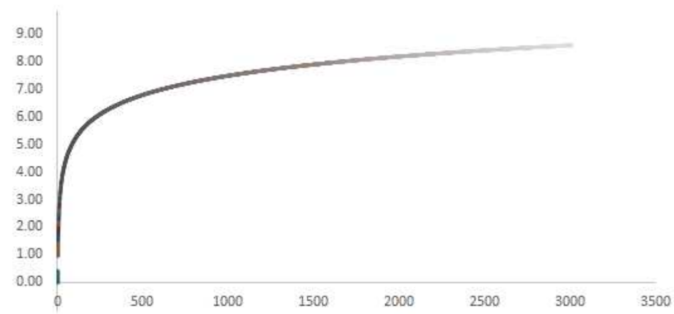
$$\sum_{k=1}^{\infty}\frac{1}{k}=\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+...+\frac{1}{k}+...\,,$$

is called the *harmonic series.*

Example 5.5.5: harmonic series

Below we show the underlying sequence, $a_n = 1/n$, that is known to converge to zero (left), and its sequence of the partial sums (right):





Plotting the first 3000 terms of the latter seems to suggests that it also converges. Examining the data, we find that the sum isn't large, so far:

$$\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+...+\frac{1}{3000}\approx 8.59\,.$$

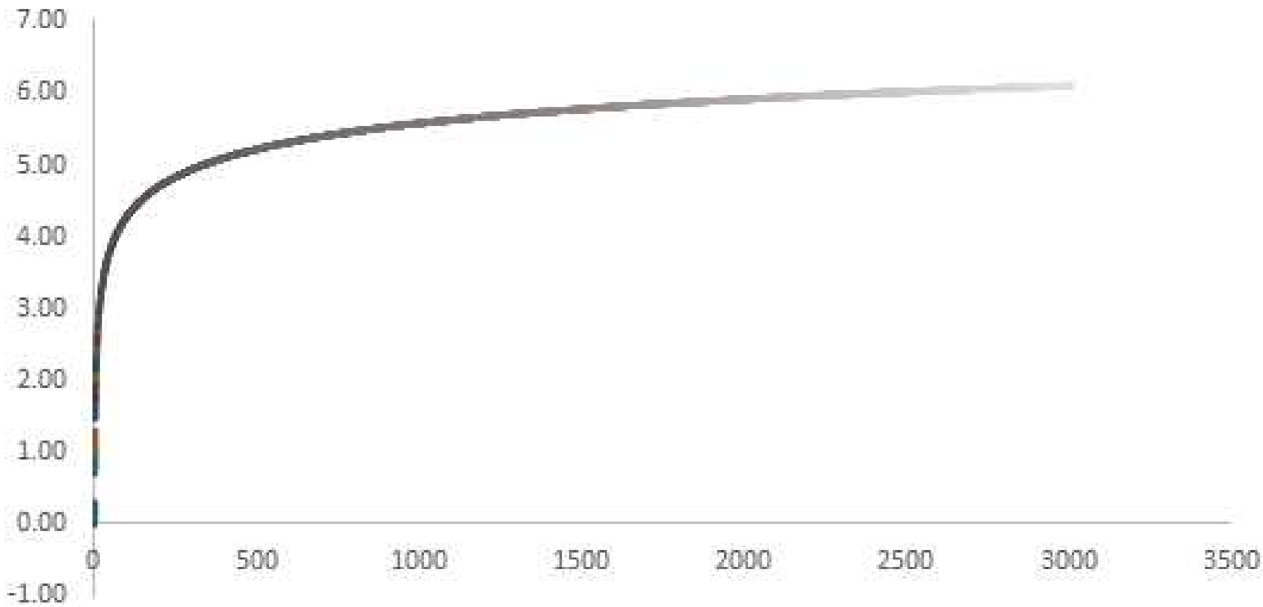
We know better than to think that this tells us anything; the series *diverges* as we will show in this chapter.

Example 5.5.6: higher powers

If we replace the power of k in the harmonic series, which is -1, with -1.1, the series,

$$\sum_{k=1}^{\infty}\frac{1}{k^{1.1}}\,,$$

the graph of its sums looks almost exactly the same:



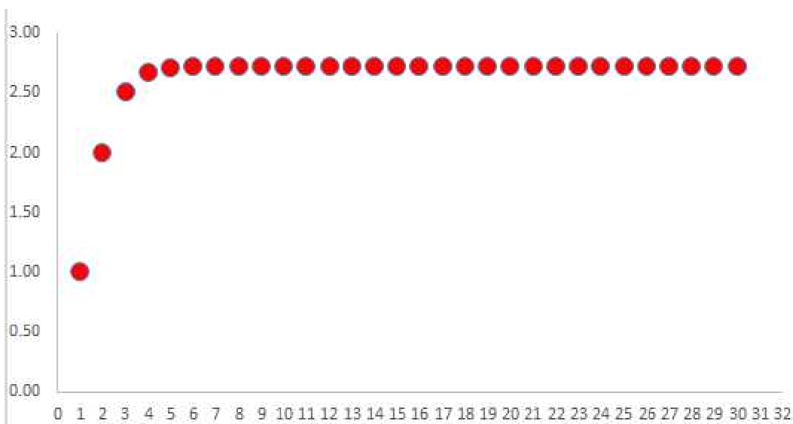
However, we show below that it is convergent!

Example 5.5.7: factorials

In contrast, this is how fast the series of the *reciprocals of the factorials*,

$$\sum_{k=1}^{\infty} \frac{1}{k!}\,,$$

converges:



As shown in earlier in this chapter, it converges to e.

One of the most important series is the following.

Definition 5.5.8: geometric series

The series produced by the geometric progression $a_n = a \cdot r^n$ with ratio r,

$$\sum_{k=0}^{\infty} ar^k = a + ar^1 + ar^2 + ar^3 + ar^4 + ... + ar^k + ...\,,$$

is called the *geometric series* with ratio r.

Recall the fact about geometric *progressions* from Volume 1 (Chapter 1DC-1).

Theorem 5.5.9: Convergence of Geometric Progression

The geometric progression with ratio r converges and diverges depending on the value of r:

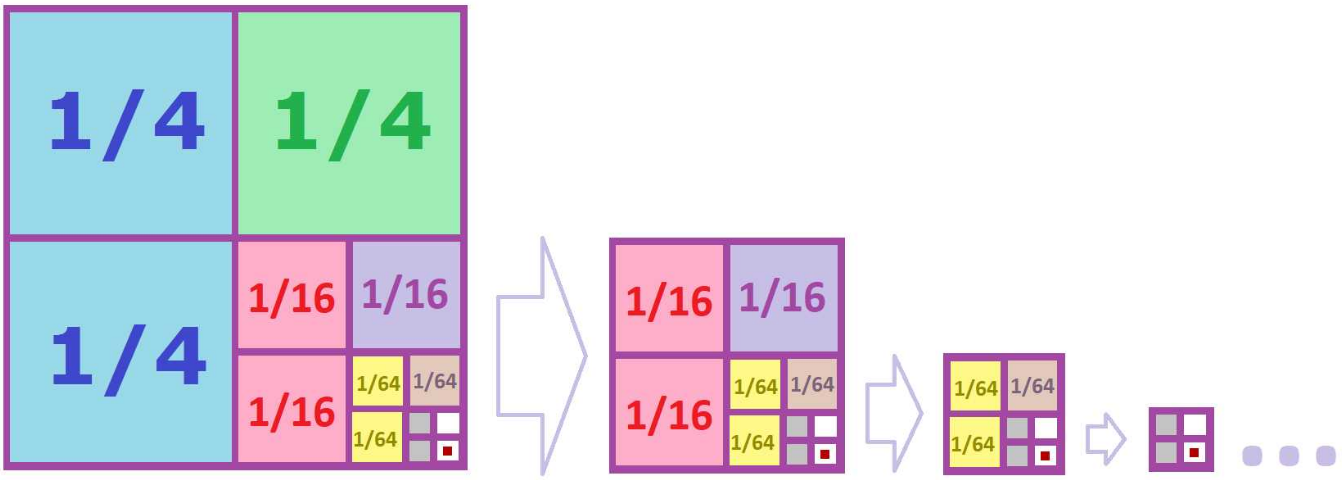
$$\lim_{n\to\infty} r^n = \begin{cases} \text{diverges} & \text{if } r \le -1, \\ 0 & \text{if } |r| < 1, \\ 1 & \text{if } r = 1, \\ +\infty & \text{if } r > 1. \end{cases}$$

Warning!

The result tells us nothing about the convergence of the series.

Example 5.5.10: geometric

The construction shown below is recursive; it cuts a square into four, takes the first three, and then repeats the procedure with the last one:



These are the areas:

first two squares	third square	first two squares	third square	...	
$\frac{1}{4}+\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{16}+\frac{1}{16}$	$\frac{1}{16}$	...	
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	...	
$\frac{1}{2}$	$+\frac{1}{4}$	$+\frac{1}{8}$	$+\frac{1}{16}$	$+\ldots$	$= 1$?

Because they are all cut out of the big square and all parts of the square are covered, the total sum must be 1. Provided such a sum makes sense in the first place! Does it? Each step creates two terms in the following series:

$$\sum_{k=1}^{n} \frac{1}{2^k} = \left(\frac{1}{2}+\frac{1}{4}\right)+\left(\frac{1}{8}+\frac{1}{16}\right)+\ldots \stackrel{?}{=} 1$$

The sum of the series then must be 1, if it converges. This is a geometric series and the construction suggests that a geometric series with ratio $r = 1/2$ converges.

Let's investigate the partial sums of the general geometric series,

$$ar^0,\ ar^1,\ ar^2,\ \ldots ar^{n-1},\ ar^n,\ \ldots,$$

under the following restrictions:

$$r \neq 0, \ a \neq 0 .$$

We will need an *explicit* formula for the sequence of sums of this sequence:

$$p_n = ar^0 + ar^1 + ar^2 + ... + ar^{n-1} + ar^n .$$

The "..." part is what stands in the way.

The following clever trick (presented first in Chapter 1PC-1) solves the problem. Below, we write the nth partial sum p_n in the first row, its multiple rp_n (all terms are multiplied by r) in the second, subtract them, and then cancel the terms that appear twice:

$$\begin{array}{lllllllr}
p_n & = ar^0 & +ar^1 & +ar^2 & +... & +ar^{n-1} & +ar^n & \text{Subtract} \\
rp_n & = \quad ar^1 & + \quad ar^2 & + \quad ar^3 & +... & + \quad ar^n & + \quad ar^{n+1} & \\
\hline
p_n - rp_n & = ar^0 - ar^1 & +ar^1 - ar^2 & +ar^2 - ar^3 & +... & +ar^{n-1} - ar^n & +ar^n - ar^{n+1} & \\
 & = ar^0 & & & & & -ar^{n+1}. &
\end{array}$$

The "..." part is gone! Therefore,

$$p_n(1-r) = a - ar^{n+1} .$$

Thus, we have an explicit formula for the nth term of the partial sum:

$$p_n = a\frac{1-r^{n+1}}{1-r} .$$

Let's note that this is the sum of a geometric progression and a constant sequence:

$$p_n = \frac{-ra}{1-r}r^n + \frac{a}{1-r} .$$

We then have an interesting match:

The sequence of sums	The integral
of a geometric progression	of an exponential function
is a geometric progression	is an exponential function
with the same ratio	with the same base
plus a constant.	plus a constant.

It's easy to find the limit of this sequence because r^{n+1} is the only term that matters. The conclusion is the following:

Theorem 5.5.11: Sum of Geometric Series

The geometric series with ratio r converges if and only if

$$|r| < 1 ;$$

in that case, the sum is the following:

$$\boxed{\sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}}$$

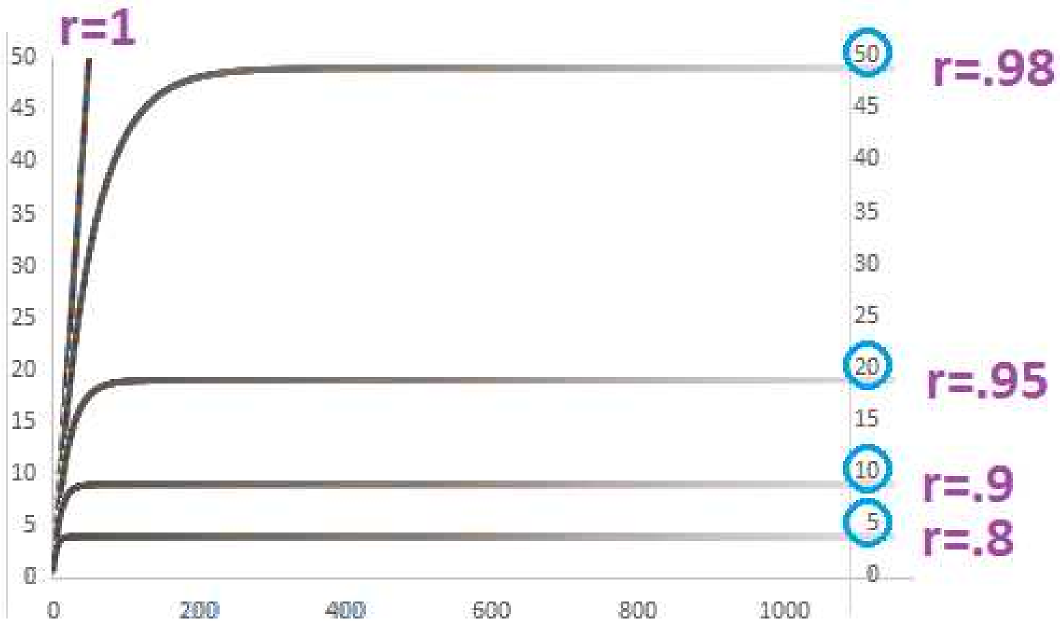
Proof.

We use the above formula and the familiar properties of limits:

$$\begin{aligned}\sum_{k=0}^{n} ar^k &= \lim_{n\to\infty} p_n\\ &= \lim_{n\to\infty} a\frac{1-r^{n+1}}{1-r}\\ &= \frac{a}{1-r}\lim_{n\to\infty}(1-r^{n+1})\\ &= \frac{a}{1-r}\left(1-\lim_{n\to\infty} r^{n+1}\right).\end{aligned}$$

To finish, we invoke the theorem above about the convergence of geometric *progressions*.

The theorem is confirmed numerically below:



Exercise 5.5.12

Explain the relation between the sequence of these geometric series and the series with $r = 1$.

Example 5.5.13: geometric series

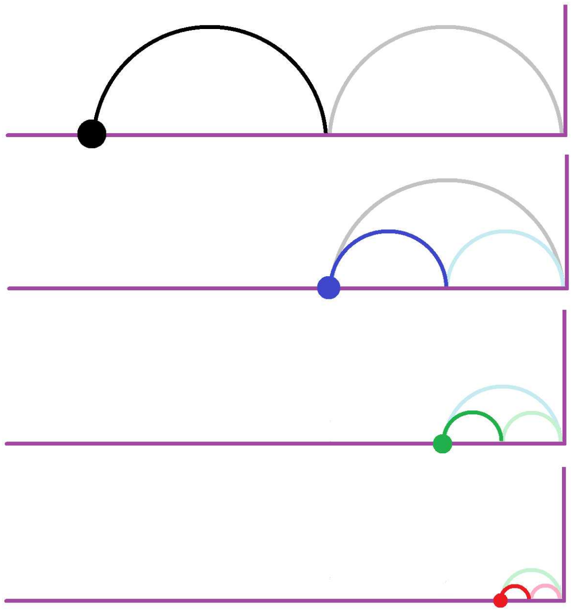
We apply the theorem as follows:

geometric series	first term a	ratio r	sum
$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+...$	1	$\frac{1}{2}$	$\frac{1}{1-1/2}=2$
$1+\frac{1}{3}+\frac{1}{9}+\frac{1}{27}+...$	1	$\frac{1}{3}$	$\frac{1}{1-1/3}=\frac{3}{2}$
$1-\frac{1}{2}+\frac{1}{4}-\frac{1}{8}+...$	1	$-\frac{1}{2}$	$\frac{1}{1+1/2}=\frac{2}{3}$
$1+1.1+1.1^2+1.1^3+...$	1	1.1	diverges
$1-1+1-1+...$	1	-1	diverges

This powerful theorem will also allow us to study other series by comparing them to a geometric series.

Example 5.5.14: Zeno's paradox

Recall a simple scenario (from Chapter 2DC-1): As you walk toward a wall, you can never reach it because once you've covered half the distance, there is still distance left, etc. It will take infinitely many steps to reach the wall:



We now know that the sum of the distances is 1 as a geometric series:

$$\frac{1}{1}+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+...=1\,.$$

But we knew that! What resolves the paradox is the fact that the *time* periods also form a geometric series. Suppose v is the speed of the person. Then, time periods are computed as the distance over the speed:

$$\frac{1}{2}/v,\ \frac{1}{4}/v,\ \frac{1}{8}/v,\ ...$$

This is a geometric sequence with:

$$a=\frac{1}{2v},\ r=1/2\,.$$

Its sum is found according to the theorem:

$$\frac{1}{2}\frac{1}{v}+\frac{1}{4}\frac{1}{v}+\frac{1}{8}\frac{1}{v}+...=\frac{1}{v}\,.$$

This is how long it takes. So, even though it takes infinitely many *steps* to reach the wall, the *time* it takes isn't infinite!

Example 5.5.15: infinite decimals

Let's represent the number 0.44444... as a series. If we can demonstrate convergence, this is what we have:

$$\begin{array}{llllll} 0.4444... & =0.4 & +0.04 & +0.004 & +0.0004 & +... \\ & =4\cdot 0.1 & +4\cdot 0.04 & +4\cdot 0.004 & +4\cdot 0.0004 & +... \\ & =4\cdot 10^{-1} & +4\cdot 10^{-2} & +4\cdot 10^{-3} & +4\cdot 10^{-4} & +... \end{array}$$

This is a geometric series with the first term $a=.4$ and the ratio $r=0.1<1$. Therefore, it converges by the theorem to the following:

$$\sum_{i=0}^{n}0.4\cdot 0.1^n=\sum_{i=0}^{n}ar^n=\frac{a}{1-r}=\frac{0.4}{1-0.1}=\frac{4}{9}\,.$$

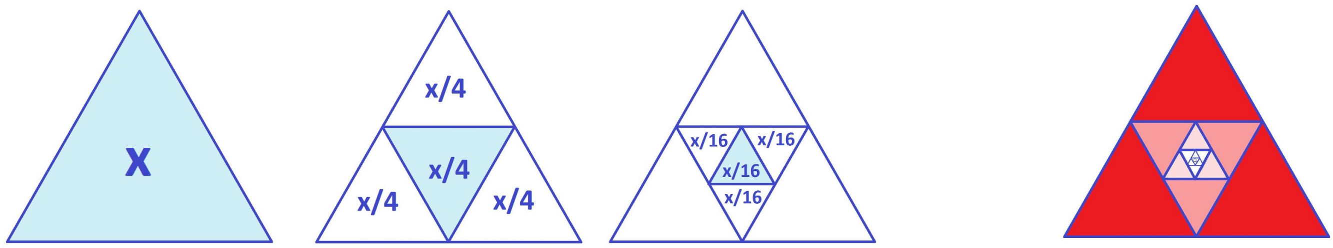
Exercise 5.5.16

Use the last example to show that for any digit d we have the following representation:

$$.\overline{dddd...} = \frac{d}{9}\,.$$

Example 5.5.17: geometric

Here is another geometric interpretation of a geometric series:



Instead of finding the sum of a series as a number, we are producing a series from a number.

We start with a simple observation:

$$1 = \frac{3}{4} + \frac{1}{4} = \frac{1}{4}(3+1)\,.$$

We replace 1 with this expression, which also contains 1, to be replaced with this expression again, etc. This way, in contrast to all other examples, we start with the sum and then acquire a series for it:

$$\begin{aligned}
1 &= \frac{1}{4}(3+1) && = \frac{3}{4} + \frac{1}{4} \\
&= \frac{1}{4}(3+1) && = \frac{3}{4} + \frac{1}{4}\cdot 1 \\
&= \frac{1}{4}\Big(3+\frac{1}{4}(3+1)\Big) && = \frac{3}{4} + \frac{1}{4}\left(\frac{3}{4} + \frac{1}{4}\right) \\
&= \frac{1}{4}\Big(3+\frac{1}{4}(3+1)\Big) && = \frac{3}{4} + \frac{3}{4^2} + \frac{1}{4^2}\cdot 1 \\
&= \frac{1}{4}\left(3+\frac{1}{4}\Big(3+\frac{1}{4}(3+1)\Big)\right) && = \frac{3}{4} + \frac{3}{4^2} + \frac{1}{4^2}\left(\frac{3}{4} + \frac{1}{4}\right) \\
&= ... \\
&= \sum_{n=1}^{\infty}\frac{3}{4^n}\,.
\end{aligned}$$

This *infinite* computation makes sense, thanks to the last theorem.

Example 5.5.18: power series

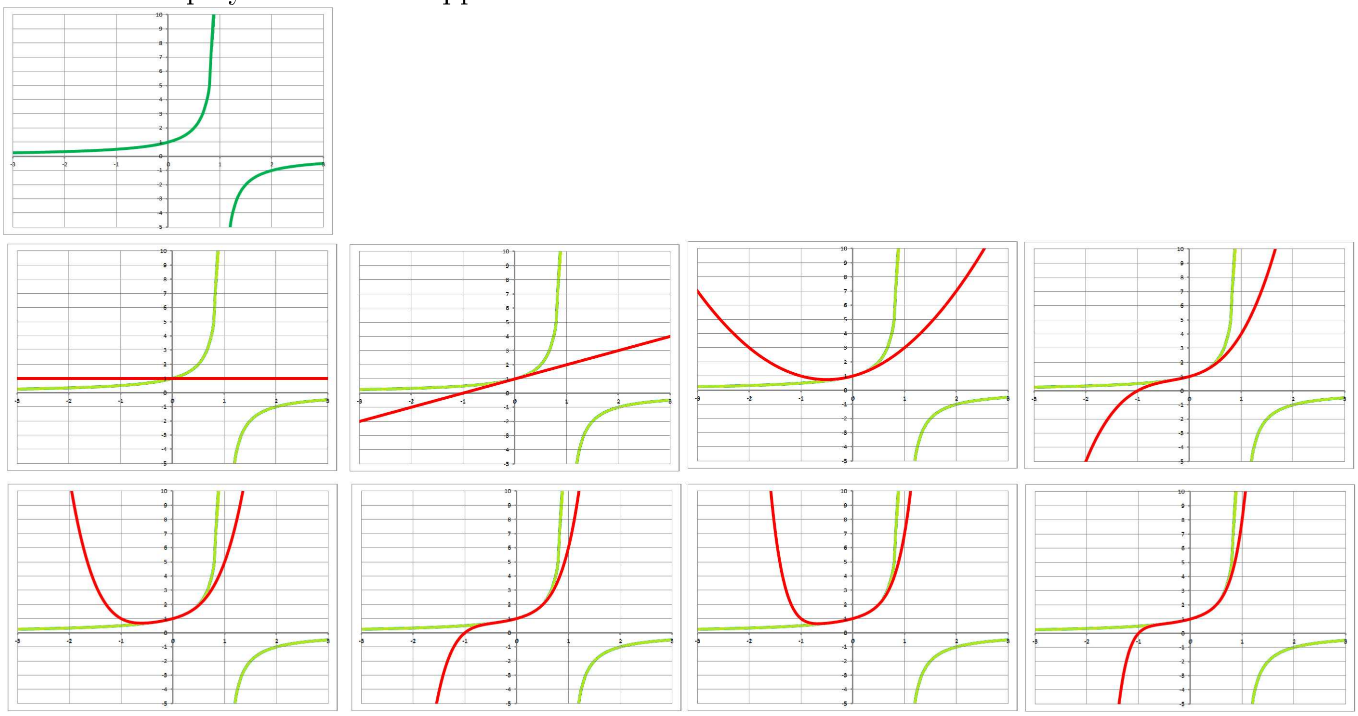
Let's not forget why we are doing this. Consider the familiar (and the simplest) power series:

$$1 + x + x^2 + x^3 + ...$$

It converges for $x=0$ and diverges for $x=2$. We know more now. For each x, this *is a geometric series* with ratio $r=x$. Therefore, it converges for every x that satisfies $|x|<1$. The sum is a number, and this number is the value of a *function* at x. The interval $(-1,1)$ is the *domain* of the function defined this way. The theorem even provides a formula for this function:

$$1 + x + x^2 + x^3 + ... = \frac{1}{1-x}\,.$$

The difference between these two functions given here is clear; it's their domains. The partial sums of the series are polynomials that approximate the function:



The mismatch is visible in the graph as the approximation fails to *improve* outside the interval $(-1, 1)$.

Exercise 5.5.19

Confirm that these are Taylor polynomials of the function.

Exercise 5.5.20

Find the sums of the series generated by each of these sequences or show that it doesn't exist:

1. $1/1,\ 1/3,\ 1/5,\ 1/7,\ 1/9,\ 1/11,\ 1/13, ...$
2. $1/0.9,\ 1/0.99,\ 1/0.999,\ 1/0.9999,\ ...$
3. $1,\ -1,\ 1,\ -1,\ ...$
4. $1,\ -1/2,\ 1/4,\ -1/8,\ ...$
5. $1,\ 1/4,\ 1/16\ , 1/64,\ ...$

Example 5.5.21: telescoping series

When a series isn't geometric, it might still be possible to simplify the partial sums via algebraic tricks and find its sum:

$$\begin{aligned}
\sum_{k=1}^{\infty} \frac{1}{k(k+1)} &= \sum_{k=1}^{\infty} \left(\frac{1}{k} - \frac{1}{k+1} \right) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} \left(\frac{1}{k} - \frac{1}{k+1} \right) \\
&= \lim_{n\to\infty} \left[\left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + ... + \left(\frac{1}{n} - \frac{1}{n+1}\right) \right] \\
&= \lim_{n\to\infty} \left[1 + \left(-\frac{1}{2} + \frac{1}{2}\right) + \left(-\frac{1}{3} + \frac{1}{3}\right) + ... + \left(-\frac{1}{n} + \frac{1}{n}\right) - \frac{1}{n+1} \right] \\
&= \lim_{n\to\infty} \left[1 - \frac{1}{n+1} \right] \\
&= 1\,.
\end{aligned}$$

Exercise 5.5.22

Explain how the above computation is different from the following and why it matters:

$$\begin{aligned}\sum_{k=1}^{\infty}\frac{1}{k(k+1)} &= \sum_{k=1}^{\infty}\left(\frac{1}{k}-\frac{1}{k+1}\right)\\ &= \left(1-\frac{1}{2}\right)+\left(\frac{1}{2}-\frac{1}{3}\right)+...+\left(\frac{1}{n}-\frac{1}{n+1}\right)+...\\ &= 1+\left(-\frac{1}{2}+\frac{1}{2}\right)+\left(-\frac{1}{3}+\frac{1}{3}\right)+...+\left(-\frac{1}{n}+\frac{1}{n}\right)+...\\ &= 1\,.\end{aligned}$$

Exercise 5.5.23

Use the algebraic trick from the above example to evaluate this integral:

$$\int\frac{1}{x(x+1)}\,dx\,.$$

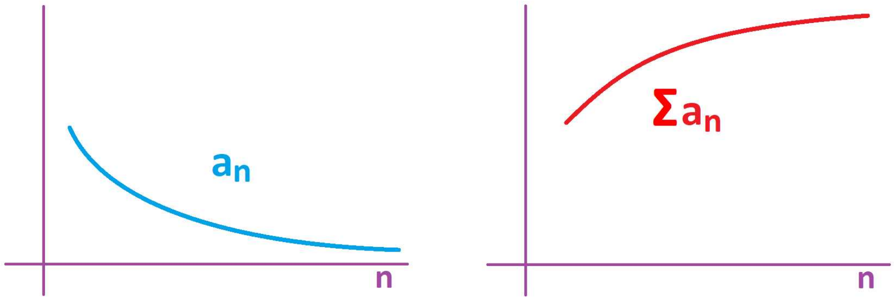
The *Geometric Series Theorem*, as well as other theorems about convergence of series presented below, allows us to be bolder with computations that involve infinitely many steps.

5.6. From finite sums via limits to series

Let's remember that, initially,

- *there are no series.*

A series is shorthand for what we do with a sequence. A series is, therefore, always a pair of sequences:



What we do is recursive and, when the sequence is infinite, involves infinitely many steps.

At the next stage, notation takes over and the series start to be treated as entities:

Series

$$\sum_{k=m}^{\infty}a_k \ \text{ is the limit of partial sums of } \ a_k\,.$$

Exercise 5.6.1

Show that every sequence is the sequence of sums of some sequence. Hint: Use the difference construction:

$$a_{n+1}=\Delta p_n=p_{n+1}-p_n\,.$$

In Volume 1 (Chapter 1PC-1) we introduced *finite sums*, i.e., sums of the terms of a given sequence a_n over an interval $[a, b]$ of values of k:

$$\sum_{k=p}^{q} a_k = a_p + a_{p+1} + \ldots + a_q \,,$$

and their properties. We used those properties to study the Riemann sums in Chapter 1; these sums are the areas under the graphs of the step-functions produced by the sequences. Now we just need to transition to *infinite sums* via limits.

As a matter of notation, we will often omit the bounds in the sigma notation for series:

Sigma notation

$$\sum_{k=m}^{\infty} a_k \text{ is replaced with } \sum a_k$$

The reason is why this is acceptable is the following. Consider these two finite sums:

$$\sum_{k=a}^{n} u_k \quad \text{and} \quad \sum_{k=b}^{n} u_k \,.$$

The difference between them is just a finitely many terms and, therefore, they either both converge or both diverge, according to the *Truncation Principle* from Volume 2 (Chapter 2DC-1).

The notation is especially appropriate when there is no hope of finding the sum of the series. In that case, we face a simple dichotomy:

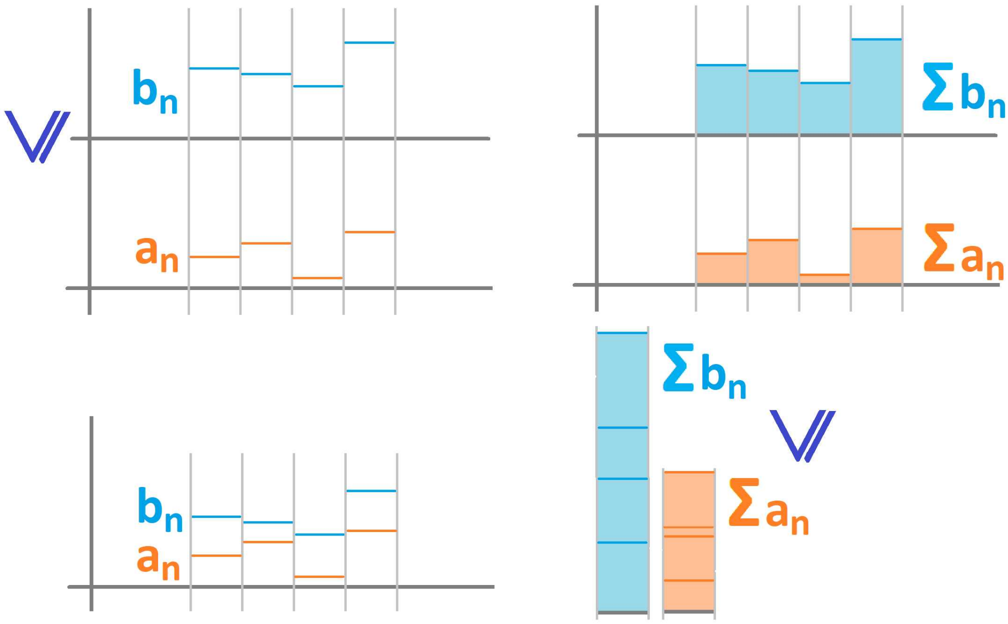
- it converges, or
- it diverges.

Next, we prove several theorems about convergence of series. They all follow the same idea:

> ▶ If an algebraic relation exists for *finite* sums, then this relation remains valid for *infinite* sums, i.e., series, provided they converge.

First, the *comparison properties.*

If two sequences are comparable, then so are their sums:



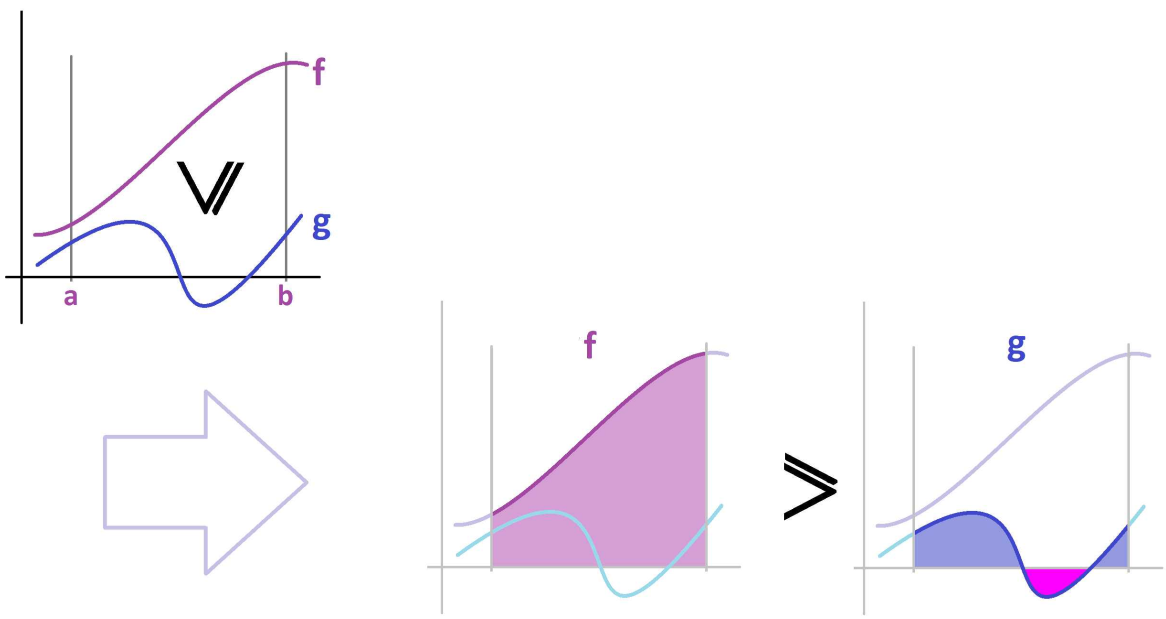
In fact, this simple algebra tells the whole story:

Add the two inequalities producing the third:
$$\begin{array}{rcl} u & \le & U \\ v & \le & V \\ \hline u+v & \le & U+V \end{array}$$

The only difference is that we have more than just two terms:

Add the sequence of inequalities producing the new one:
$$\begin{array}{rcl} u_p & \le & U_p \\ u_{p+1} & \le & U_{p+1} \\ ... & ... & ... \\ u_q & \le & U_q \\ \hline u_p + ... + u_q & \le & U_p + ... + U_q \\ \sum_{n=p}^{q} u_n & \le & \sum_{n=p}^{q} U_n \end{array}$$

This is the *Comparison Rule for Sums* presented in Volume 1 (Chapter 1PC-1) that was used to study Riemann sums and then Riemann integrals in Chapter 1. Zooming out helps us see that the larger function, or a sequence, always contains a larger area under its graph:



Taking the limit $q \to \infty$ allows us to make the following conclusion based on the *Comparison Rule for Limits of Sequences*.

Theorem 5.6.2: Comparison Rule for Series

An inequality between the corresponding terms of two series holds for their sums too.

In other words, suppose u_n and U_n are sequences. Suppose also that either of the two series they produce converges. Then:

$$u_n \le U_n \implies \sum u_n \le \sum U_n\,.$$

Example 5.6.3: comparison with geometric

The series
$$\sum_{n=1}^{\infty} \frac{1}{2^n-1}$$
isn't geometric... but close. Consider this obvious fact:
$$2^n - 1 \le 2^n \implies \frac{1}{2^n-1} \ge \frac{1}{2^n}\,.$$
The series produced by the sequence on the right converges to 1. It follows according to the theorem that
$$\sum_{n=1}^{\infty} \frac{1}{2^n-1} \ge 1\,,$$
if we can establish that this series converges.

A related result is the following.

Theorem 5.6.4: Strict Comparison Rule for Series

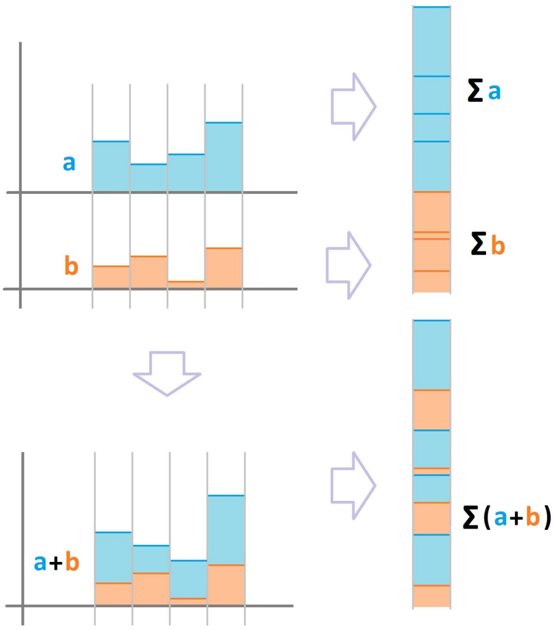
Suppose u_n and U_n are sequences. Suppose also that either of the two series they produce converges. Then:
$$u_n < U_n \implies \sum u_n < \sum U_n\,.$$

Exercise 5.6.5

Prove the theorem.

Now the *algebraic properties* of the series.

The picture below illustrates the idea of adding two series:



In fact, this simple algebra, the *Associative Property*, tells the whole story:

$$\text{Re-arrange the terms:}\quad \begin{array}{rccc} & u & + & U \\ + & & & \\ & v & + & V \\ \hline = & (u+v) & + & (U+V) \end{array}$$

The only difference is that we have more than just two terms:

Re-arrange the sum of two sequences:
$$\begin{array}{rcl} u_p & + & U_p \\ u_{p+1} & + & U_{p+1} \\ \dots & \dots & \dots \\ u_q & + & U_q \\ \hline u_p+\dots+u_q & + & U_p+\dots+U_q \\ =(u_p+U_p)+ & \dots & +(u_q+U_q) \\ =\sum_{n=p}^{q}(u_n & + & U_n) \end{array}$$

This is the *Sum Rule for Sums* we saw in Volume 1 (Chapter 1PC-1). Taking the limit $q\to\infty$ allows us to make the following conclusion based on the *Sum Rule for Limits*.

Theorem 5.6.6: Sum Rule for Series

If two convergent series are added term by term, then their sums are added too.

In other words, suppose u_n and U_n are sequences. If the series they produce converge, then so does their term-by-term sum, and we have:

$$\sum(u_n+U_n)=\sum u_n+\sum U_n\,.$$

Example 5.6.7: adding series

Compute the sum:

$$\sum_{n=1}^{\infty}\left(\frac{1}{2^n}+\frac{e^{-n}}{3}\right).$$

This sum is a limit, and we can't assume that the answer will be a number. Let's *try* to apply the *Sum Rule*:

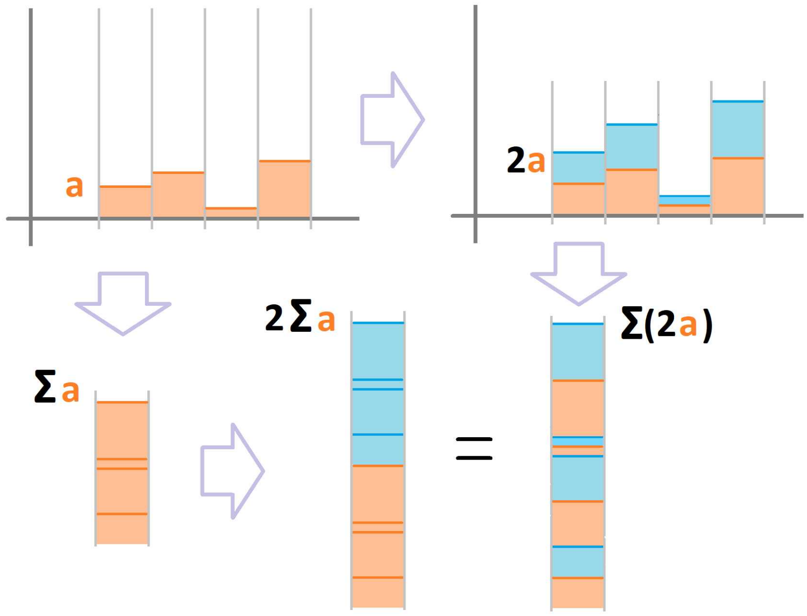
$$\begin{array}{lllll}\displaystyle\sum_{n=1}^{\infty}\left(\frac{1}{2^n}+\frac{e^{-n}}{3}\right) & \overset{?}{=}\displaystyle\sum_{n=1}^{\infty}\frac{1}{2^n} & +\displaystyle\sum_{n=1}^{\infty}\frac{e^{-n}}{3} & \text{Yes, but only if the two sums converge!}\\ & =\displaystyle\sum_{n=1}^{\infty}\left(\frac{1}{2}\right)^n & +\displaystyle\sum_{n=1}^{\infty}\frac{1}{3}\left(e^{-1}\right)^n & \text{Do they?}\end{array}$$

These are two geometric series with the ratios equal to, respectively, $1/2$ and $1/e$. Both numbers are smaller than 1 and, therefore, either series converges! This conclusion justifies our use of the Sum Rule and the computation that follows:

$$\begin{array}{llll}\displaystyle\sum_{n=1}^{\infty}\left(\frac{1}{2^n}+\frac{e^{-n}}{3}\right) & =\displaystyle\sum_{n=1}^{\infty}\left(\frac{1}{2}\right)^n & +\displaystyle\sum_{n=1}^{\infty}\frac{1}{3}\left(e^{-1}\right)^n & \\ & =\dfrac{1/2}{1-1/2} & +\dfrac{1/3}{1-1/e}\,. & \text{This is the answer!}\end{array}$$

We used the formula, $\frac{a}{1-r}$, from the last section.

The picture below illustrates the idea of multiplication of the terms of a series:



In fact, this simple algebra, the *Distributive Property*, tells the whole story:

$$\text{Factor out:}\quad \begin{array}{rccc} k\cdot(& u & + & U) \\ = & ku & + & kU \end{array}$$

The only difference is that we have more than just two terms:

$$\text{Factor out:}\quad \begin{array}{ccl} k & \cdot & u_{,} \\ k & \cdot & u_{p+1} \\ \ldots & \ldots & \ldots \\ k & \cdot & u_q \\ \hline k & \cdot & (u_p + \ldots + u_q) \\ = k & \cdot & \displaystyle\sum_{n=p}^{q} u_n \end{array}$$

This is the *Constant Multiple for Sums* we saw in Volume 1 (Chapter 1PC-1). Taking the limit $q \to \infty$ allows us to make the following conclusion based on the *Constant Multiple for Limits*.

Theorem 5.6.8: Constant Multiple Rule for Series

If a convergent series is multiplied by a number term by term, then its sum is multiplied by that number too.

In other words, suppose u_n is a sequence and k is a number. If the series converges, then so does its term-by-term multiple, and we have:

$$\sum (k \cdot u_n) = k \cdot \sum u_n\,.$$

Example 5.6.9: power series

Note that we can also add two power series term by term:

$$\begin{aligned} & \sum_{n=m}^{\infty} a_n(x-a)^n && + \sum_{n=m}^{\infty} b_n(x-a)^n \\ = & \sum_{n=m}^{\infty} \Big(a_n(x-a)^n && + \ b_n(x-a)^n \Big) \\ = & \sum_{n=m}^{\infty} (a_n+b_n)(x-a)^n, \end{aligned}$$

for every x for which the series converge, creating a new power series. Note also that we can multiply a power series term by term:

$$\begin{aligned} & k \cdot \sum_{n=m}^{\infty} a_n(x-a)^n \\ = & \sum_{n=m}^{\infty} k \cdot a_n(x-a)^n \\ = & \sum_{n=m}^{\infty} (ka_n)(x-a)^n, \end{aligned}$$

for every x for which the series converges, creating a new power series. They behave just like *polynomials*!

Just as with integrals, there are no direct counterparts for the Product Rule and the Quotient Rule.

Just as with integrals, these rules are shortcuts: They allow us to avoid having to use the definition every time we need to compute a sum.

Let's make sure that our *decimal system* is on a solid ground. What is the meaning of

$$\pi = 3.14159...?$$

The infinite decimal is understood as an infinite sum:

$$\begin{array}{llllllll} 3.14159... & = 3 & +0.1 & +0.04 & +0.001 & +0.0005 & +0.00009 & +... \\ & = 3 & +1\cdot 0.1 & +4\cdot 0.01 & +1\cdot 0.001 & +5\cdot 0.0001 & +9\cdot 0.00001 & +... \\ & = 3 & +1\cdot 0.1^1 & +4\cdot 0.1^2 & +1\cdot 0.1^3 & +5\cdot 0.1^4 & +9\cdot 0.1^5 & +... \end{array}$$

The power indicates the placement of the digit within the representation. In general, we have a sequence of integers between 0 and 9, i.e., *digits*, $d_1, d_2, ..., d_n, ...$. It is meant to define a real number via a series:

$$\overline{.d_1 d_2 ... d_n ...} \stackrel{?}{=} \sum_{k=1}^{\infty} d_k \cdot (0.1)^k \,.$$

The theorem proves that such a definition makes sense.

Theorem 5.6.10: Decimals as Series

Suppose d_k is a sequence of integers between 0 and 9. Then the series

$$\sum_{k=1}^{\infty} d_k \cdot 0.1^k = d_1 \cdot 0.1 + d_2 \cdot 0.01 + d_3 \cdot 0.001 + ...$$

converges.

Proof.

First, the sequence of partial sums of this series is *increasing*:

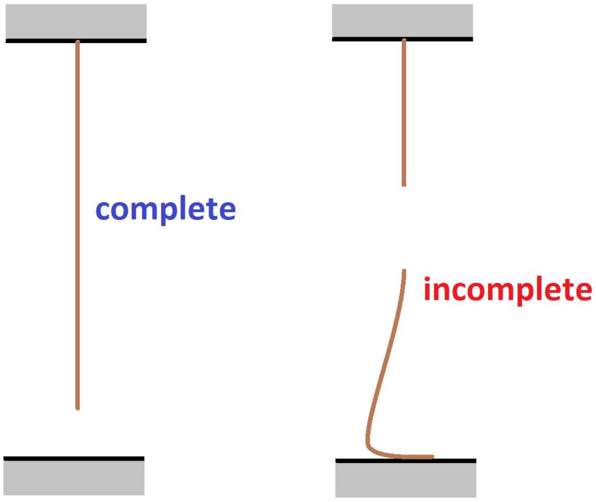
$$p_{n+1} = p_n + d_{n+1} \cdot 0.1^{n+1} > p_n \,.$$

It is also *bounded*:

$$\begin{aligned} p_n &= d_1 \cdot 0.1 + d_2 \cdot 0.01 + d_3 \cdot 0.001 + ... + d_n \cdot 0.1^n \\ &< 10 \cdot 0.1 + 10 \cdot 0.01 + 10 \cdot 0.001 + ... + 10 \cdot 0.1^n \\ &< 10 \cdot 0.1 + 10 \cdot 0.01 + 10 \cdot 0.001 + ... + 10 \cdot 0.1^n + ... \\ &= \frac{1}{1-0.1} \qquad \text{Because it's a geometric series with } r = 0.1\,. \\ &= 9\,. \end{aligned}$$

Therefore, the sequence is convergent by the *Monotone Convergence Theorem* (Volume 2).

The result explains why the Monotone Convergence Theorem is also known as the *Completeness Property of Real Numbers*; an "incomplete" rope won't hang:



Exercise 5.6.11

Prove the theorem by using the Comparison Rule for Series instead.

Exercise 5.6.12

State and prove an analogous theorem for *binary* arithmetic.

Moreover, the decimal numbers are also subject to the algebraic operations according to the algebraic theorems above.

Example 5.6.13: algebra of decimals

We can add infinite decimals:

$$\begin{array}{llll} u & = \overline{.u_1u_2...u_n...} & = \sum_{k=1}^{\infty} u_k \cdot (0.1)^k \\ v & = \overline{.v_1v_2...v_n...} & = \sum_{k=1}^{\infty} v_k \cdot (0.1)^k \\ \hline u+v & & = \sum_{k=1}^{\infty} (u_k+v_k) \cdot (0.1)^k \end{array}$$

We can also multiply an infinite decimal by another real number:

$$\begin{aligned} u &= \overline{.u_1u_2...u_n...} &&= \sum_{k=1}^{\infty} u_k \cdot (0.1)^k \\ c \cdot u & &&= \sum_{k=1}^{\infty} c \cdot u_k \cdot (0.1)^k \end{aligned}$$

The formulas don't tell us the decimal representations of $u+v$ (unless $u_n + v_n < 10$) or $c \cdot u$.

5.7. Divergence

What if we face algebraic operations with series that *diverge*? The laws above tell us nothing! For example, this is the *Sum Rule*:

$$\sum a_n,\ \sum b_n \text{ converge } \Longrightarrow \sum (a_n + b_n) \text{ converges;}$$

but this isn't:

$$\sum a_n,\ \sum b_n \text{ diverge } \not\Longrightarrow \sum (a_n + b_n) \text{ diverges.}$$

Exercise 5.7.1

Show that this statement would indeed be untrue.

Example 5.7.2: convergent plus divergent

Compute the sum:

$$\sum_{n=1}^{\infty} \left(\frac{1}{2^n} + \frac{e^n}{3} \right).$$

This sum is a limit, and we can't assume that the answer will be a number. Let's *try* to apply the Sum Rule:

$$\begin{aligned} \sum_{n=1}^{\infty} \left(\frac{1}{2^n} + \frac{e^n}{3} \right) &\overset{?}{=} \sum_{n=1}^{\infty} \frac{1}{2^n} &&+ \sum_{n=1}^{\infty} \frac{e^n}{3} && \text{Yes, but only if the two series converge!} \\ &= \sum_{n=1}^{\infty} \left(\frac{1}{2} \right)^n &&+ \sum_{n=1}^{\infty} \frac{1}{3} e^n . && \text{Do they?} \end{aligned}$$

These are two geometric series with the ratios equal to, respectively, $1/2$ and e. The first one is smaller than 1 and therefore the series converges, but the second one is larger than 1 and therefore the series diverges! The conclusion is that it is *unjustified to use the Sum Rule*. We pause at this point... and then try to solve the problem by other means.

We recall from Chapter 2DC-1 that, in the case of infinite limits, we adhere to the following rules ($k \neq 0$):

Algebra of Infinities

$$\begin{array}{lcll} \text{number} & + & (\pm\infty) & = \pm\infty \\ \pm\infty & + & (\pm\infty) & = \pm\infty \\ k & \cdot & (\pm\infty) & = \pm\operatorname{sign}(k)\infty \end{array}$$

It follows that the series in the last example diverges to infinity.

These formulas suggest the following divergence result for series.

Theorem 5.7.3: Push Out Theorem for Series

1. *If the values of a series lie above those of a series that diverges to positive infinity, then so does this series.*
2. *If the values of a series lie below those of a series that diverges to negative infinity, then so does this series.*

In other words, suppose a_n and b_n are sequences. Suppose that, for an integer p, we have:

$$a_n \geq b_n \quad \text{for } n \geq p\,.$$

Then:

$$\begin{array}{ccc}\sum a_n = +\infty & \Longleftarrow & \sum b_n = +\infty \\ \sum a_n = -\infty & \Longrightarrow & \sum b_n = -\infty\end{array}$$

Proof.

It follows from the *Push Out Theorem for Sequences* (Volume 2).

So, the smaller series, if it goes to $+\infty$, pushes the larger one up, to $+\infty$. And the larger series, if it goes to $-\infty$, pushes the larger one down, to $-\infty$.

Example 5.7.4: comparison

Consider this obvious fact:

$$\frac{1}{2-1/n} \geq \frac{1}{2}\,.$$

It follows that

$$\sum_{n=1}^{\infty} \frac{1}{2-1/n} = \infty\,.$$

Exercise 5.7.5

Give examples of series that show that the converse of the theorem is untrue.

Warning!

Not all divergent series diverge to infinity.

We turn to algebra:

Theorem 5.7.6: Divergence of Sum of Series

Suppose a_n and b_n are sequences. Then, we have:

- *The sum of a divergent series and a convergent series diverges:*

$$\sum a_n \text{ diverges}, \quad \sum b_n \text{ converges} \quad \Longrightarrow \quad \sum (a_n + b_n) \text{ diverges.}$$

- *For series with non-negative terms, the sum of two divergent series diverges:*

$$\sum a_n \text{ diverges}, \quad \sum b_n \text{ diverges} \quad \Longrightarrow \quad \sum (a_n + b_n) \text{ diverges.}$$

Theorem 5.7.7: Divergence of Constant Multiple of Series

Suppose a_n is a sequence. Then, a constant, non-zero multiple of a divergent series diverges:

$$\sum a_n \text{ diverges } \implies \sum ka_n \text{ diverges, } k \neq 0\,.$$

Exercise 5.7.8

Prove these theorems.

The pattern that we may have noticed is that constantly adding positive numbers that grow will give you infinity at the limit. The same conclusion is, of course, true for a constant sequence. What if the sequence decreases? Then it depends. For example, $a_n = 1 + 1/n$ decreases but the series still diverges. It appears that the sequence should at least decrease to zero.

The actual result is crucial.

Theorem 5.7.9: Divergence Test for Series

If a sequence doesn't converge to zero, its sum diverges.

In other words, we have:

$$\boxed{\sum a_n \text{ converges } \implies a_n \to 0}$$

Proof.

We have to invoke the definition. But let's turn to the *contrapositive form* of the theorem:

▶ If a series converges, then the underlying sequence converges to 0.

In other words, we need to prove this:

$$\lim_{n\to\infty} \sum_{k=1}^{n} a_k = P \implies \lim_{n\to\infty} a_n = 0\,,$$

where P is some number. Or, for the sequence of partial sums p_n, we have to demonstrate this:

$$\lim_{n\to\infty} p_n = P \implies \lim_{n\to\infty} a_n = 0\,.$$

Recall the recursive formulas for the sequence of sums of a_n:

$$p_{n+1} = p_n + a_n\,,$$

and, accordingly, we have the original as the sequence of differences of p_n:

$$a_n = p_n - p_{n-1}\,.$$

Therefore,

$$\begin{aligned}
\lim_{n\to\infty} a_n &= \lim_{n\to\infty} (p_n - p_{n-1}) && \text{Both converge, and we apply...}\\
&= \lim_{n\to\infty} p_n - \lim_{n\to\infty} p_{n-1} && \text{...the Sum Rule for Sequences.}\\
&= P - P && \text{The limit is the same because it's the same sequence.}\\
&= 0\,.
\end{aligned}$$

The converse of the theorem is untrue as will be seen from the example of the harmonic series:

$$\sum a_n \quad \text{converges} \quad \Longrightarrow a_n \to 0$$
$$\sum a_n \quad \text{converges} \quad \not\Longleftarrow a_n \to 0$$

Example 5.7.10: test divergence

The test is for *divergence* and nothing else:

$$\begin{array}{llll} 1. & \lim \left(1+\frac{1}{n}\right) \neq 0 & \Longrightarrow & \sum \left(1+\frac{1}{n}\right) \text{ diverges.} \\ 2. & \lim \sin n \neq 0 & \Longrightarrow & \sum \sin n \text{ diverges.} \\ 3. & \lim \frac{1}{n} = 0 & \Longrightarrow & \text{test fails.} \\ 4. & \lim \frac{1}{n^2} = 0 & \Longrightarrow & \text{test fails.} \end{array}$$

Warning!

The failure of the test doesn't prove anything!

5.8. *Series with non-negative terms*

The convergence of such series is easier to determine.

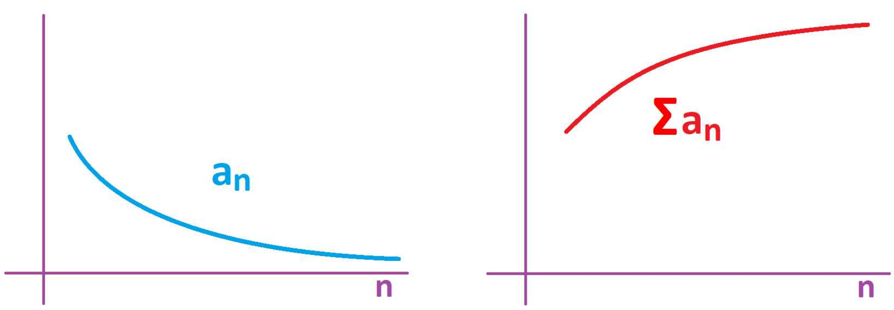
All we need is the *Monotone Convergence Theorem* from Volume 2 (Chapter 2DC-1):

- ▶ Every monotone bounded sequence converges.

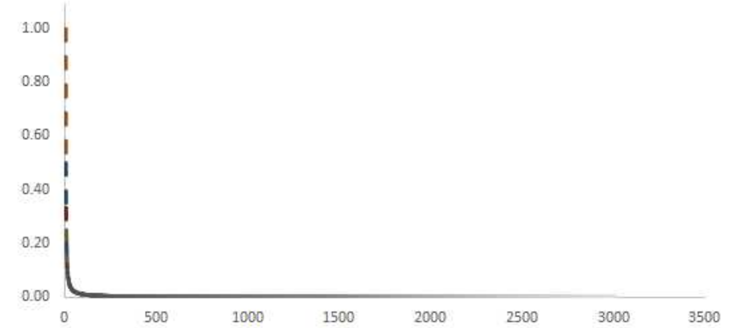
Of course, we will not apply the theorem to the sequence but rather to its sequence of sums! Indeed, if the original sequence a_n has non-negative terms: $a_n > 0$, then the new sequence of sums is *increasing*:

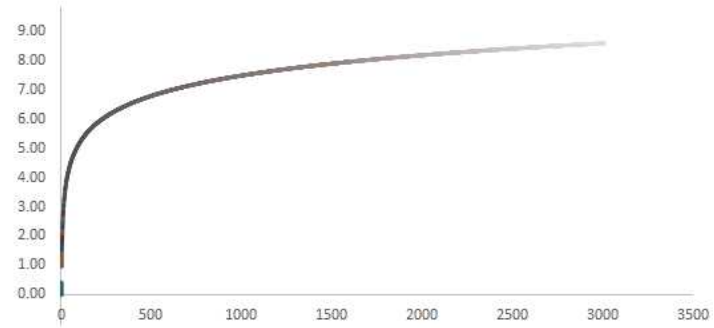
$$p_{n+1} = p_n + a_{n+1} \geq p_n \,.$$

If the *Divergence Test* is also satisfied, the pair of the sequence and its sums will have to look just like the generic illustration we have been using:



A good example is the harmonic series below:





So, if such a sequence is also *bounded*, it is convergent. Therefore, such a series can't *just* diverge, as $\sum \sin n$ does; it has to *diverge to infinity*.

Theorem 5.8.1: Non-negative Series

For a series $\sum a_n$ with non-negative terms, $a_n \geq 0$, there can be only have two outcomes:
- *it converges, or*
- *it diverges to infinity.*

This observation significantly simplifies things; the sum of the series is either:

- a number, or
- the infinity.

The latter option and the former option are often written in the following notation whenever the value of the sum is not being considered:

Sum of series with non-negative terms

$$\sum a_n = \infty \quad \text{or} \quad \sum a_n < \infty$$

Many theorems in the rest of the chapter will only tell the former from the latter...

Example 5.8.2: divergence

Since

$$\lim \left(1 + \frac{(-1)^n}{n}\right) = 1 \neq 0\,,$$

the series

$$\sum \left(1 + \frac{(-1)^n}{n}\right)\,,$$

fails the *Divergence Test* and, therefore, diverges. Moreover,

$$\sum \left(1 + \frac{(-1)^n}{n}\right) = \infty\,.$$

Exercise 5.8.3

Prove that the harmonic series diverges by following this observation: for each k consecutive terms, $\frac{1}{k+1}$, $\frac{1}{k+2}$, ..., $\frac{1}{2k}$, they are all $\geq \frac{1}{2k}$, so their sum is $\geq \frac{k}{2k} = \frac{1}{2}$.

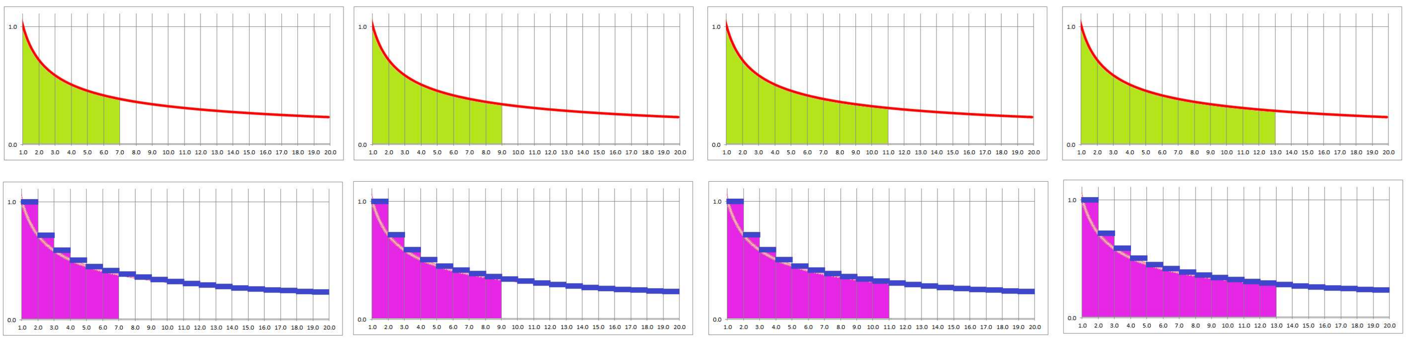
We now address the issue of series vs. improper integrals over infinite domains.

Both are limits:

$$\begin{aligned}\int_1^\infty f(x)\,dx &= \lim_{b\to\infty}\int_1^b f(x)\,dx\\ \sum_{i=1}^\infty a_i &= \lim_{n\to\infty}\sum_{i=1}^n a_i\end{aligned}$$

And the notation matches too.

Not only the integral but also the sum on the right represent areas under graphs of certain functions defined on finite intervals. Both the integral and the sum on the left are computed as limits of the ones on the right:

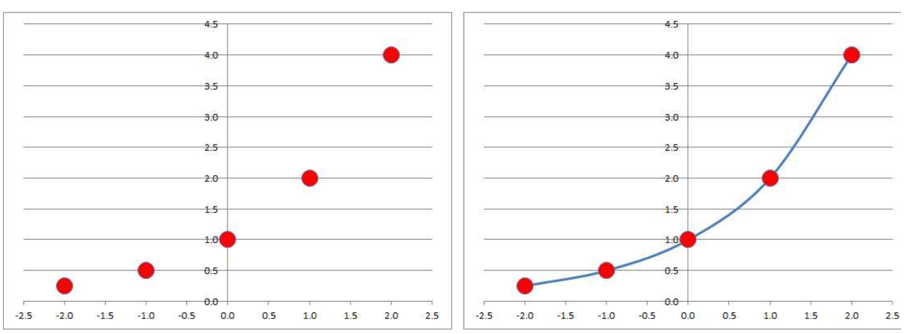


We can conjecture now that if f and a_n are related, then these limits, though not equal, may be related too.

The sequence may come from *sampling* the function:

$$a_n = f(n),\ n = 1, 2, 3, \ldots$$

Like this:

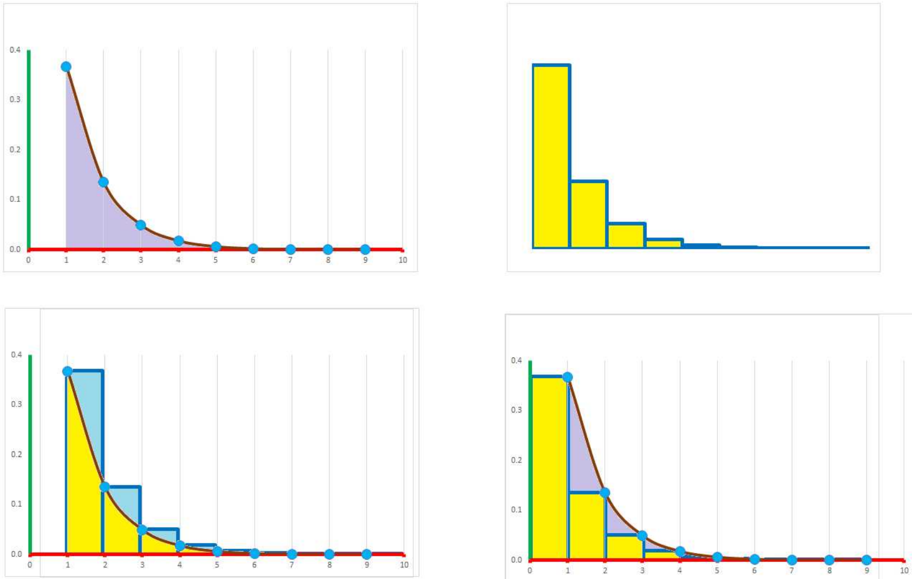


Example 5.8.4: integrals vs. sums

Consider this pair:

$$f(x) = e^{-x} \ \text{ and } \ a_n = e^{-n}\,.$$

What is the relation between *the integral of the former and the sum of the latter*? Both are the areas of certain regions and we can place one below or above the other:



The improper integral is easy to compute:

$$\int_1^\infty e^{-x}\,dx = \lim_{b\to\infty}\int_1^b e^{-x}\,dx = \lim_{b\to\infty} -(e^{-b} - e^1) = e\,.$$

Meanwhile, the series is geometric with $r = 1/e$:

$$\sum_{n=1}^\infty e^{-n} = \frac{1/e}{1-1/e}\,.$$

Both converge!

A more general result is the following. They either both converge or both diverge (to infinity) for any $a > 0$:

- the improper integral of the exponential function with base $1/a$ over $[1,\infty)$ and
- the geometric series with ratio $1/a$.

In other words, we have:

$$\int_1^\infty a^{-x}\,dx < \infty \iff \sum_{n=1}^\infty a^{-n} < \infty\,.$$

Exercise 5.8.5

Prove the statement.

We will now try to apply the idea to a general series with non-negative terms if we can match it with an improper integral. The hope is that we can handle integrals better – with all the tools in Chapter 2 – than the series.

First, we might discover that our series is "dominated" by a *convergent* improper integral:

$$\sum_{k=1}^\infty a_n \le \int_1^\infty f(x)\,dx < \infty\,.$$

The condition is satisfied when:

$$a_n \le f(x) \ \text{ for every } x \text{ in } [n, n+1]\,.$$

Or, we might discover that our series "dominates" a *divergent* improper integral:

$$\sum_{k=1}^{\infty} a_n \geq \int_1^{\infty} f(x)\, dx = \infty\,.$$

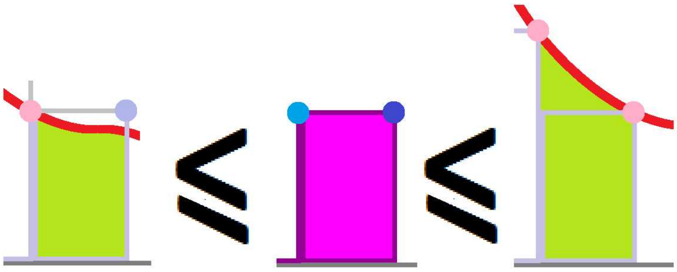
The condition is satisfied when:

$$a_n \geq f(x) \quad \text{for every } x \text{ in } [n, n+1]\,.$$

There is a way to combine these two conditions into one. The idea is to "squeeze" the sequence between *two* functions. But where does the other function come from? We shift the graph of f to the right by 1 unit in order to put it *above* the sequence:

$$f(x) \leq a_n \leq f(x-1)\,.$$

This is the main idea:



The conditions that make sure that the picture is justified are listed below.

Theorem 5.8.6: Integral Comparison Test

Suppose that on $[1, \infty)$,

1. f *is continuous;*
2. f *is decreasing;*
3. f *is non-negative.*

Suppose also that we have a sequence:

$$a_n = f(n),\ n = 1, 2, \ldots$$

Then the improper integral and the series below,

$$\int_1^{\infty} f(x)\, dx \quad \text{and} \quad \sum_{n=1}^{\infty} a_n\,,$$

either both converge or both diverge (to infinity); i.e.,

$$\int_1^{\infty} f(x)\, dx < \infty \iff \sum_{n=1}^{\infty} a_n < \infty\,.$$

Proof.

From the second condition, we conclude that for every $n = 2, 3, \ldots$ and all $n \leq x \leq n+1$, we have

$$f(x) \leq a_n \leq f(x-1)\,.$$

Therefore, we have:

$$\int_n^{n+1} f(x)\, dx \leq a_n \leq \int_n^{n+1} f(x-1)\, dx\,,$$

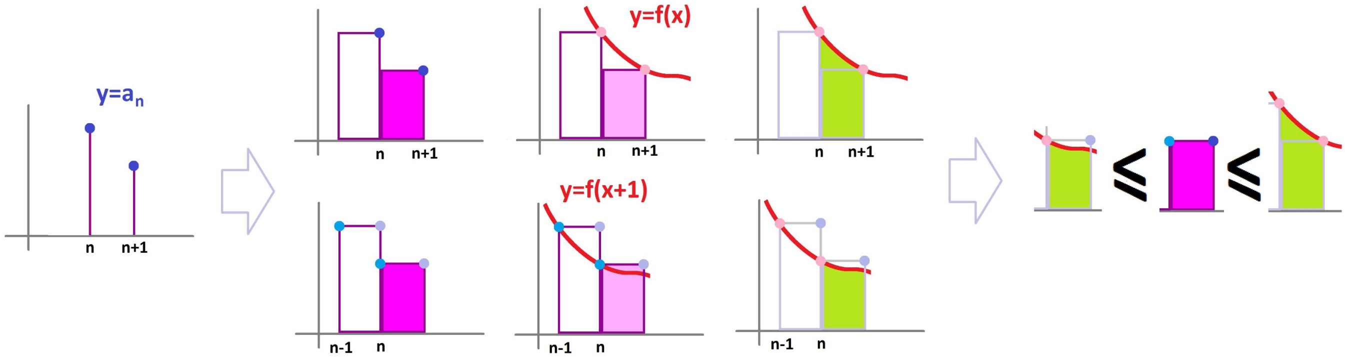
or, after a substitution on the right,

$$\int_n^{n+1} f(x)\,dx \le a_n \le \int_{n-1}^{n} f(x)\,dx\,.$$

Adding all these for $n = 2, 3, 4, \ldots$, we obtain:

$$\int_2^{\infty} f(x)\,dx \le \sum_{n=2}^{\infty} a_n \le \int_1^{\infty} f(x)\,dx\,.$$

Now, either of the two conclusions of this theorem follows from the corresponding part of the *Non-negative Series Theorem* above. This is the main step of the proof:



Exercise 5.8.7

Demonstrate that none of the three conditions of the theorem can be dropped.

Warning!

The sum of the series can be estimated by the integrals but remains unknown.

Example 5.8.8: harmonic series diverges

Unlike a geometric series, the sums of the harmonic series doesn't have an explicit formula. A comparison is, therefore, necessary. The harmonic series diverges because

$$\int_1^{\infty} \frac{1}{x}\,dx = \lim_{x\to+\infty} \ln x = +\infty\,.$$

More general is the following.

Corollary 5.8.9: p-series

A p-series, i.e.,

$$\sum \frac{1}{n^p}\,,$$

- *converges when $p > 1$ and*
- *diverges when $0 < p \le 1$.*

Proof.

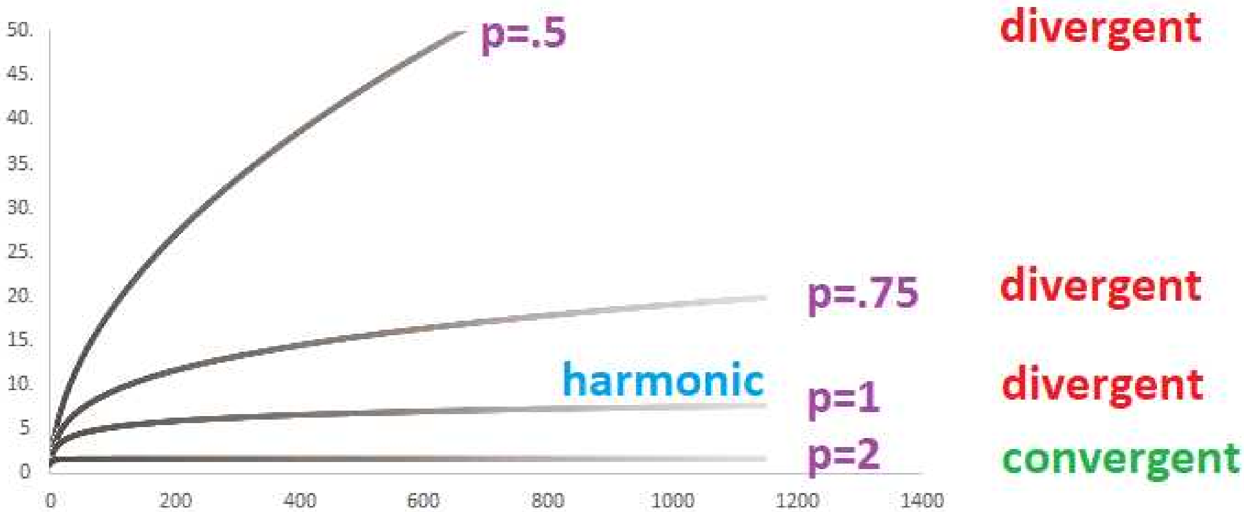
Once the function f is chosen:

$$f(x) = \frac{1}{x^p}\,,$$

the rest of the proof is purely computational. Indeed,

$$\begin{aligned}
\int_1^\infty f(x)\,dx &= \int_1^\infty x^{-p}\,dx \\
&= \begin{cases} \lim\limits_{b\to\infty} \dfrac{1}{-p+1} x^{-p+1}\Big|_1^b & \text{if } p \neq 1, \\ \lim\limits_{b\to\infty} \ln x \Big|_1^b & \text{if } p = 1, \end{cases} \\
&= \begin{cases} \lim\limits_{b\to\infty} \dfrac{1}{-p+1} (b^{-p+1} - 1^{-p+1}) & \text{if } p \neq 1, \\ \lim\limits_{b\to\infty} (\ln b - \ln 1) \Big|_1^b & \text{if } p = 1, \end{cases} \\
&= \begin{cases} \dfrac{1}{-p+1} & \text{if } p > 1, \\ \infty & \text{if } p = 1, \quad \text{The harmonic series.} \\ \infty & \text{if } p < 1. \end{cases}
\end{aligned}$$

Thus, not only the harmonic series diverge but it also separates the divergent p-series from the convergent ones:



Exercise 5.8.10

Show that the theorem fails if we drop the assumption that the function is decreasing. Can this assumption be weakened?

Exercise 5.8.11

Show that the theorem fails if we drop the assumption that the function is non-negative. Can this assumption be weakened?

5.9. Comparison of series

In the last section, we matched series with improper integrals in order to derive the convergence or divergence of the former from that of the latter. Now we follow this idea but, instead, compare series to other series.

Warning!

The Comparison Rule for Series doesn't help here because it assumes the convergence of both series.

The plan is as follows:

- Compare a new series (with non-negative terms) to an old one the convergence/divergence of which is known.

The starting point is the *Comparison Rule for Sums* (Chapter 1PC-1): The sum of a sequence with smaller elements is smaller. In other words, if $a_n \le b_n$, then we have for any p, q with $p \le q$:

$$\sum_{n=p}^{q} a_n \le \sum_{n=p}^{q} b_n \,.$$

Therefore, the sequence of partial sums of the first series is "dominated" by that of the second. If the second converges, its sum is an upper bound for the partial sums of the first. Then, the first series converges too by the theorem in the last section.

This is the main result.

Theorem 5.9.1: Direct Comparison Test for Series

Suppose we have two series with non-negative terms that satisfy the following:

$$0 \le a_n \le b_n \,,$$

for all n. Then, then the convergence/divergence of the larger/smaller implies the convergence/divergence of the smaller/larger; i.e.,

$$\begin{array}{ccc} \sum a_n < \infty & \Longleftarrow & \sum b_n < \infty \\ \sum a_n = \infty & \Longrightarrow & \sum b_n = \infty \end{array}$$

Exercise 5.9.2

Prove the second part.

In order to apply the theorem to a particular – new – series, we should try to modify its formula while paying attention to whether it is getting smaller or larger.

Example 5.9.3: p-series as a start

Let's go backwards at first. We consider the *p-series* and see whether we can derive the known convergence facts about them.

Some series can be *modified into* such a series. The one on the left is unfamiliar but can be matched

with a familiar one:
$$\frac{1}{n^2+1} \le \frac{1}{n^2}\,.$$
We remove "+1" and the new series is "larger". Then, by the theorem, the convergence of the original series (on the left) follows from the convergence of this p-series with $p = 2 > 1$:
$$\sum \frac{1}{n^2+1} \le \sum \frac{1}{n^2} < \infty\,.$$

Similarly, we can modify this series:
$$\frac{1}{n^{1/2}-1} \ge \frac{1}{n^{1/2}}\,.$$
We remove "-1" and the new series is "smaller". Then, the divergence of the original series (on the left) follows from the divergence of this p-series with $p = 1/2 < 1$:
$$\sum \frac{1}{n^{1/2}+1} \ge \sum \frac{1}{n^{1/2}} = \infty\,.$$

Now let's try to modify this series:
$$\frac{1}{n^2-1} \ge \frac{1}{n^2}\,.$$
We remove "-1" and the new series is "smaller". Then, the divergence of the original series (on the left) follows from the divergence... wait, this p-series with $p = 2 > 1$ converges! So, we have:
$$\sum \frac{1}{n^2-1} \ge \sum \frac{1}{n^2} < \infty\,.$$
There is nothing we can conclude from this observation.

Similarly, let's try to modify this series:
$$\frac{1}{n^{1/2}+1} \le \frac{1}{n^{1/2}}\,.$$
We remove "+1" and the new series is "larger". Then, the convergence of the original series (on the left) follows from the convergence... but this p-series with $p = 1/2 < 1$ diverges! So, we have:
$$\sum \frac{1}{n^{1/2}-1} \le \sum \frac{1}{n^{1/2}} = \infty\,.$$
There is nothing we can conclude from this observation.

Example 5.9.4: p-series as a goal

Removing "-1" and "+1" failed to produce *useful* series for these two:
$$\frac{1}{n^2-1} \quad\text{and}\quad \frac{1}{n^{1/2}+1}\,.$$
We will have to be subtler in finding comparisons.

Let's try multiplication. We add "2" in the numerator and discover that the following is true for all $n = 2, 3, \ldots$:
$$\frac{1}{n^2-1} \le \frac{2}{n^2}\,.$$
The new series is "larger". Then, the convergence of the original series (on the left) follows from the

convergence of this (multiple of a) p-series with $p = 2 > 1$:

$$\sum \frac{1}{n^2 - 1} \le \sum \frac{1}{n^2} < \infty \,.$$

Next, we follow this idea for the other series. We add "2" in the denominator and discover that the following is true for all $n = 1, 2, ...$:

$$\frac{1}{n^{1/2} + 1} \ge \frac{1}{2n^{1/2}} \,.$$

The new series is "smaller". Then, the divergence of the original series (on the left) follows from the divergence of this (multiple of a) p-series with $p = 1/2 < 1$:

$$\sum \frac{1}{n^{1/2} + 1} \ge \sum \frac{1}{2n^{1/2}} = \infty \,.$$

What is the lesson? It is hard to compare sequences in this manner.

But did we compare *functions* previously? A way to compare the behavior of of two functions f, g at infinity is to consider their relative *order of magnitude* as presented in Chapter 2DC-6. It is defined via the *limit* of their ratio:

$$\lim_{x \to +\infty} \frac{f(x)}{g(x)} = L \,.$$

If L is infinite, we say that f has a greater magnitude:

$$f >> g \,.$$

In particular, we have the following hierarchy:

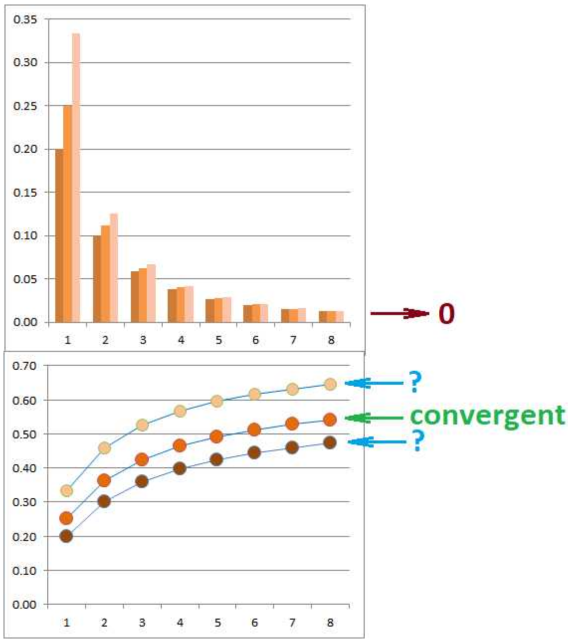
$$... >> e^x >> ... >> x^n >> ... >> x^2 >> x >> \sqrt{x} >> ... >> 1 >> ... >> \frac{1}{x} >> \frac{1}{x^2} >> ... >> e^{-x} >> ...$$

We apply this idea to series.

Example 5.9.5: comparison

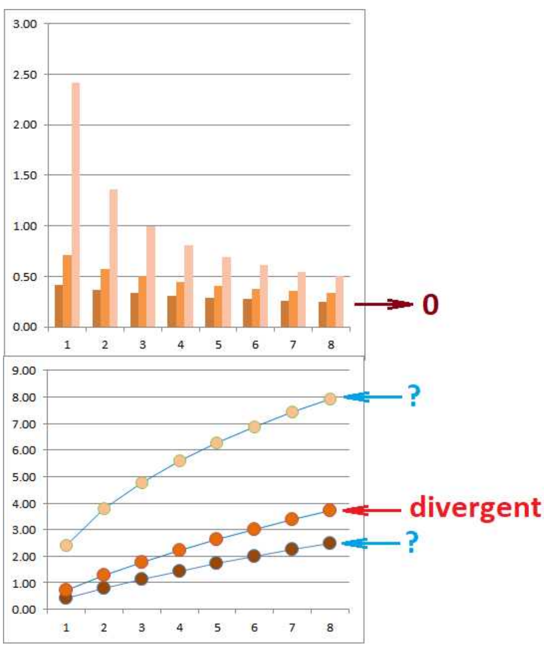
The idea is that some pairs of series converge or diverge together when they are comparable in some way. If one of them is familiar, we may have discovered what the others do. Here are the first three to be compared:

$$\frac{1}{n^2 - 1} \quad \text{vs.} \quad \frac{1}{n^2} \quad \text{vs.} \quad \frac{1}{n^2 + 1} \,.$$



This is the second triple:

$$\frac{1}{n^{1/2}-1} \quad \text{vs.} \quad \frac{1}{n^{1/2}} \quad \text{vs.} \quad \frac{1}{n^{1/2}+1}\,.$$



We consider these ratios:

$$\frac{1}{n^2 \pm 1} \div \frac{1}{n^2} \to 1 \quad \text{and} \quad \frac{1}{n^{1/2} \pm 1} \div \frac{1}{n^{1/2}} \to 1\,.$$

The second series converges in the former case and diverges in the latter. Then so does the first one.

We summarize this idea below.

Definition 5.9.6: order of magnitude for sequences

If the limit below is infinite, or its reciprocal is zero,

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \infty \quad \text{or} \quad \lim_{n\to\infty} \frac{b_n}{a_n} = 0\,,$$

and both sequences have non-negative terms, we say that a_n is of *higher order*

than b_n, and we use the following notation:

$$\boxed{a_n >> b_n \text{ and } b_n = o(a_n)}$$

The latter reads "little o". When

$$a_n >> b_n \text{ and } b_n >> a_n\,,$$

we say that they have *the same magnitude*, and we use the following notation:

$$\boxed{a_n \sim b_n}$$

We apply this terminology to both sequences and series.

Theorem 5.9.7: Limit Comparison Test

Suppose we have two sequences with non-negative terms. Then, the convergence/divergence of the larger/smaller in magnitude series implies the convergence/divergence of the smaller/larger series; i.e., there are three cases for $a_n \geq 0,\ b_n \geq 0$*:*

$$\begin{array}{lllll} \textit{Case 1,} & a_n \sim b_n & \sum a_n < \infty & \iff & \sum b_n < \infty\,. \\ \textit{Case 2,} & a_n << b_n & \sum b_n < \infty & \implies & \sum a_n < \infty\,. \\ \textit{Case 3,} & a_n >> b_n & \sum b_n = \infty & \implies & \sum a_n = \infty\,. \end{array}$$

Proof.

Suppose the limit of their ratio below exists, as a number or as infinity:

$$\lim_{n\to\infty} \frac{a_n}{b_n} = L\,.$$

In Cases 1 and 2, L is a number. Then the *definition of the limit of a sequence* (Chapter 2DC-1) states:

- For each $\varepsilon > 0$ there is such an N that for every $n > N$ we have:

$$\left| \frac{a_n}{b_n} - L \right| < \varepsilon\,.$$

Let's choose $\varepsilon = 1$. Then, for the found N, we have

$$\frac{a_n}{b_n} < L + \varepsilon = L + 1\,,$$

for every $n < N$. In other words, we have a comparison of (the tails of) two sequences:

$$a_n < (L+1)b_n\,.$$

Now, by the *Constant Multiple Rule for Series*, we have:

$$\sum b_n < \infty \implies \sum (L+1)b_n < \infty\,.$$

Then by the *Direct Comparison Test*, we have:

$$\sum a_n < \infty\,.$$

Exercise 5.9.8

Prove Case 3.

In other words, if
$$\lim_{n\to\infty}\frac{a_n}{b_n}=L$$
is a number or infinity, then these are the three cases:

$$\begin{array}{lll}\text{Case 1,} & L>0: & \text{A “perfect” match: both converge or both diverge.}\\ \text{Case 2,} & L=0: & \text{The denominator “dominates” the numerator.}\\ \text{Case 3,} & L=\infty: & \text{The numerator “dominates” the denominator.}\end{array}$$

Exercise 5.9.9

Apply the theorem to the two triples of series in the examples above.

Example 5.9.10: find a comparison series

Consider the series:
$$\sum\frac{1}{\sqrt{n^2+n+1}}\,.$$
We need to determine to what simpler series this series is “similar”. The leading term of the expression inside the radical is n^2. Therefore, we should compare our series to the following:
$$\sum\frac{1}{\sqrt{n^2}}=\sum\frac{1}{n}\,,$$
the divergent harmonic series. We evaluate the limit of the ratio now:
$$\begin{aligned}\frac{1}{\sqrt{n^2+n+1}}\div\frac{1}{n}\ &=\frac{n}{\sqrt{n^2+n+1}}\\ &=\frac{1}{\sqrt{n^2+n+1}/n}\\ &=\frac{1}{\sqrt{(n^2+n+1)/n^2}}\\ &=\frac{1}{\sqrt{1+1/n+1/n^2}}\\ &\to\frac{1}{\sqrt{1+0+0}} \qquad\qquad \text{as } n\to\infty\\ &=1\,.\end{aligned}$$

So, Case 1 of the *Limit Comparison Theorem* applies and we conclude that our series diverges.

Exercise 5.9.11

Justify the intermediate steps in the above computation.

Exercise 5.9.12

How does the theorem apply if we remove $+n$ from the above series?

Exercise 5.9.13

How does the theorem apply if we replace n^2 with n^3 in the above series?

Example 5.9.14: power series

The idea of the theorem applies well to *power series*. If we have these two,

$$\sum c_n(x-a)^n \quad \text{and} \quad \sum d_n(x-a)^n\,,$$

the condition of the theorem becomes:

$$\frac{c_n(x-a)^n}{d_n(x-a)^n} = \frac{c_n}{d_n} \to L\,,$$

when $c_n, d_n > 0$ and $x > a$. However, are these series with non-negative terms? No; just passing through $x = a$ will change the sign of the term for each odd n! The applicability of these theorems is very limited.

5.10. Absolute convergence

From this point, we abandon the severe restriction that the terms of the series can't be negative. Then, none of the results in the last two sections applies!

The plan is to make from our series a non-negative series and see if this new series converges or diverges in hope that this will tell us something about the original series.

Of course, the easiest way to do this is to take the absolute value:

$$\sum a_n \quad \text{creates} \quad \sum |a_n|\,.$$

We can also walk in the opposite direction and take any series with non-negative terms, $b_n \geq 0$, and make it "alternate":

$$\sum b_n \quad \text{creates} \quad \sum (-1)^n b_n,\ b_n \geq 0\,.$$

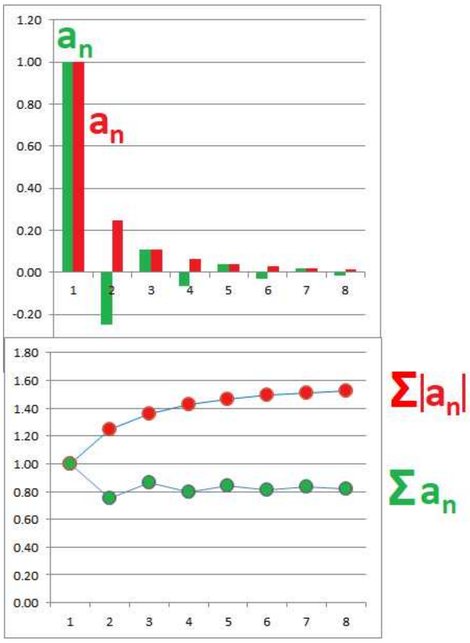
Now, which is *more* likely likely to converge, the former or the latter?

Example 5.10.1: familiar series

Some of those are easy to analyze:

- $\sum (-1)^n$ diverges – according to the *Divergence Test*.
- $\sum 1$ diverges too.
- $\sum (-1)^n \frac{1}{2^n}$ converges – according to the *Geometric Series Theorem*.
- $\sum \frac{1}{2^n}$ converges too.

In general, this is what such a pair of series looks like (the sequences are above and their sequences of partial sums are below):



As we know and can see here, the non-negative one produces the sequence of sums that is increasing. It may or may not converge depending on how much we add at every step. But for the latter, half of these up-steps are *canceled* by the down-steps! This suggests that if the former is slowing down, then so is the latter. The convergence of the former then implies the convergence of the latter...

Example 5.10.2: p-series

Consider the familiar the p-series with $p=2$:

$$\sum \frac{1}{n^2}\,.$$

It is convergent. Its "alternating" version is

$$\sum (-1)^n \frac{1}{n^2}\,.$$

They are shown above. How do these two *compare*? The former appears "smaller" than the latter:

$$(-1)^n \frac{1}{n^2} \le \frac{1}{n^2}\,.$$

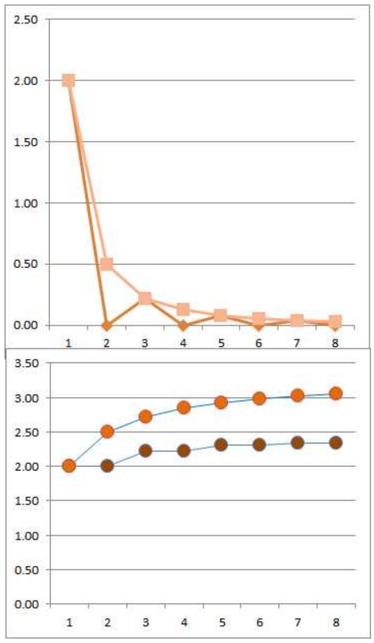
Does it mean that it must converge? No, it might still diverge because the *Non-negative Series Theorem* doesn't apply. However, a clever trick is to use a *hidden* non-negative series; it is the sum of the two:

$$(-1)^n \frac{1}{n^2} + \frac{1}{n^2} \ge 0\,.$$

What do we know about it? An insightful observation is the following inequality:

$$(-1)^n \frac{1}{n^2} + \frac{1}{n^2} \le \frac{2}{n^2}\,,$$

We have two non-negative-term series and the bigger one is convergent!



Therefore, the smaller series is convergent too,

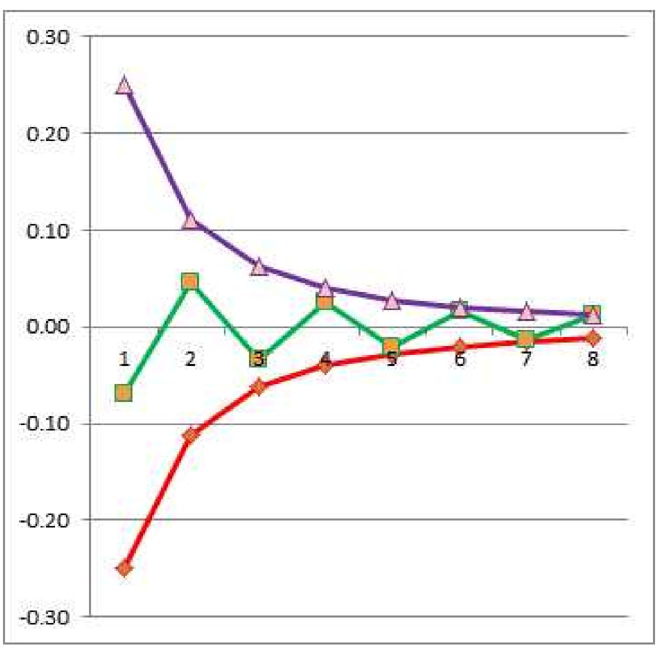
$$\sum \left((-1)^n \frac{1}{n^2} + \frac{1}{n^2} \right) < \infty \,.$$

Therefore, so is the original series,

$$\sum (-1)^n \frac{1}{n^2} \,,$$

by the *Sum Rules for Series*.

Let's generalize this example. We are after a particular kind of squeeze:



Theorem 5.10.3: Squeeze Theorem for Series

Suppose sequences $a_n,\ b_n$ with $b_n \geq 0$ satisfy:

$$-b_n \leq a_n \leq b_n$$

for all n. Then, if $\sum b_n$ converges, then so does $\sum a_n$.

Proof.

First, note that the squeeze that we have proves nothing about a_n unless we have $b_n \to 0$. Even then, all we derive is that $a_n \to 0$ too. What about the *series*? We take a different approach. First, we add b_n to the three parts of the above inequality:

$$0 \le a_n + b_n \le b_n + b_n = 2b_n \,.$$

Let's define a new sequence:

$$c_n = a_n + b_n$$

for all n. Then, we have a *new squeeze*:

$$0 \le c_n \le 2b_n \,.$$

The last sequence produces a series, $\sum 2b_n$, that converges by the *Constant Multiple Rule for Series*. Then $\sum c_n$ also converges by the *Non-negative Series Theorem* and the *Direct Comparison Theorem*. Finally, the original series,

$$\sum a_n = \sum (c_n - b_n) \,,$$

converges as the difference of two convergent series by the *Sum Rules for Series*.

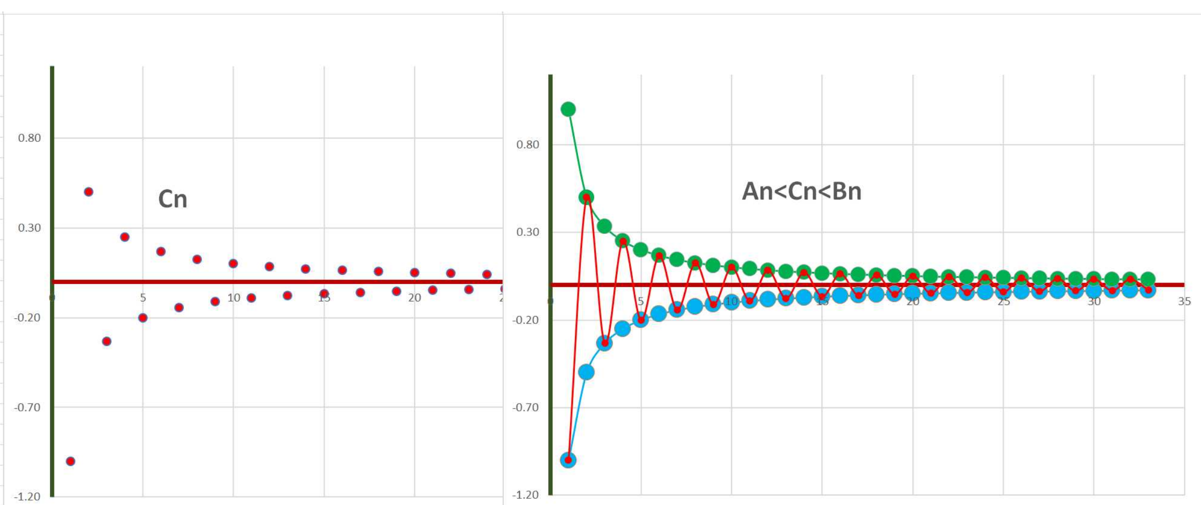
So, we can use the fact that b_n converges to prove that a_n converges too. But how do we find this convenient b_n?

There is one natural choice:

$$b_n = |a_n| \,.$$

This time the squeeze is "perfect": Not only the sequence is bounded by those two; it is, in fact, always equal to one or the other! Here is an illustration:

n	An	Cn	Bn
1	-1.00	-1.00	1.00
2	-0.50	0.50	0.50
3	-0.33	-0.33	0.33
4	-0.25	0.25	0.25
5	-0.20	-0.20	0.20
6	-0.17	0.17	0.17
7	-0.14	-0.14	0.14
8	-0.13	0.13	0.13
9	-0.11	-0.11	0.11
10	-0.10	0.10	0.10
11	-0.09	-0.09	0.09
12	-0.08	0.08	0.08
13	-0.08	-0.08	0.08
14	-0.07	0.07	0.07
15	-0.07	-0.07	0.07
16	-0.06	0.06	0.06
17	-0.06	-0.06	0.06
18	-0.06	0.06	0.06
19	-0.05	-0.05	0.05
20	-0.05	0.05	0.05
21	-0.05	-0.05	0.05
22	-0.05	0.05	0.05



It's as if a ball is continuously bouncing off the ceiling and the floor of a corridor...

The following concept is crucial:

Definition 5.10.4: series converges absolutely

We say that a series $\sum a_n$ *converges absolutely* if its series of absolute values, $\sum |a_n|$, converges.

Warning!

The word "absolute" refers to the absolute value.

The last theorem implies the following important result.

Theorem 5.10.5: Absolute Convergence Theorem

If a series converges absolutely, then it converges; i.e.,

$$\boxed{\sum |a_n| < \infty \implies \sum a_n \text{ converges}}$$

Exercise 5.10.6

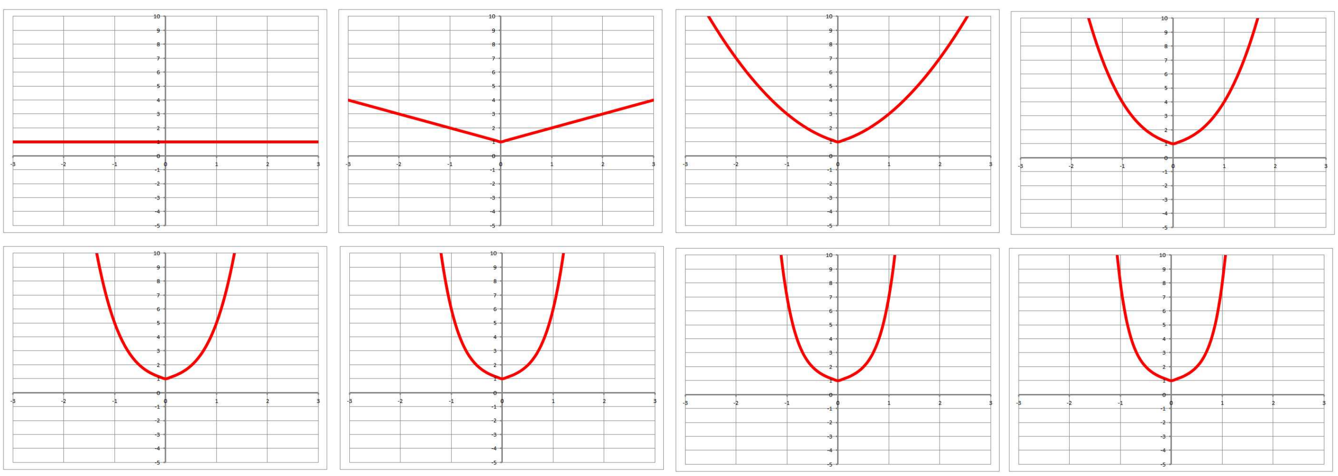
(a) Show that if a series has only positive terms, then absolute convergence means convergence. (b) Show that if a series has only negative terms, then absolute convergence means convergence.

Example 5.10.7: power series

For power series, the theorem becomes:

$$\sum |c_n||x-a|^n < \infty \implies \sum c_n(x-a)^n \text{ converges}\,,$$

for each x. The series of absolute values of the Taylor series of the exponential function is illustrated below:



Look at the cusps; these aren't powers and the new series is *not* a power series.

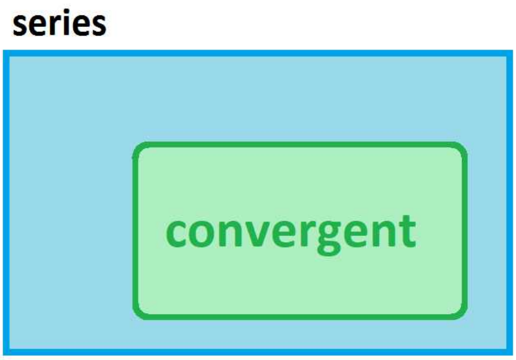
The converse of the theorem is false:

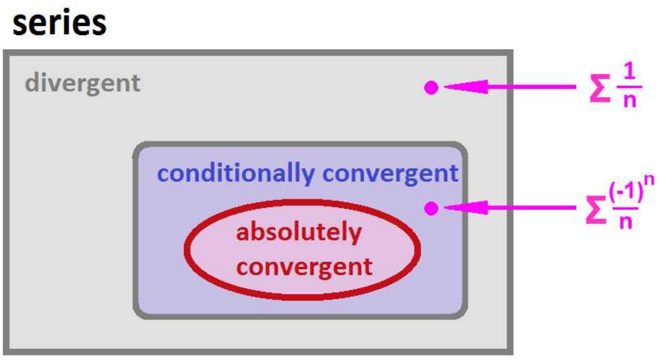
$$\begin{aligned} \sum |a_n| < \infty &\implies \sum a_n \text{ converges} \\ \sum |a_n| < \infty &\not\Longleftarrow \sum a_n \text{ converges} \end{aligned}$$

Definition 5.10.8: series converges conditionally

We say that a series $\sum a_n$ *converges conditionally* if it converges but its series of absolute values, $\sum |a_n|$, does not.

Then, every (numerical) series converges either absolutely or conditionally.





Example 5.10.9: non-negative

Of course, all convergent non-negative-term series converge absolutely. For example, all p-series with $p > 1$,

$$\sum \frac{1}{n^p}\,,$$

converge and, therefore, converge absolutely.

Conversely, for every convergent series with non-negative terms, we now know other (absolutely) convergent series. For example, since a p-series with $p > 1$,

$$\sum \frac{1}{n^p}\,,$$

converges, then its alternating version,

$$\sum \frac{(-1)^n}{n^p}\,,$$

is also convergent, absolutely, because:

$$\left|\frac{(-1)^n}{n^p}\right| = \frac{1}{n^p}\,.$$

Example 5.10.10: p-series

On the other hand, as a p-series with $p \le 1$,

$$\sum \frac{1}{n^p}\,,$$

diverges, does it mean that its alternating version,

$$\sum \frac{(-1)^n}{n^p}\,,$$

also diverges? No.

The following resolves the issue.

Theorem 5.10.11: Leibniz Alternating Series Test

Suppose a sequence b_n satisfies:

1. $b_n > 0$ *for all* n;
2. $b_n > b_{n+1}$ *for all* n; *and*
3. $b_n \to 0$ *as* $n \to \infty$.

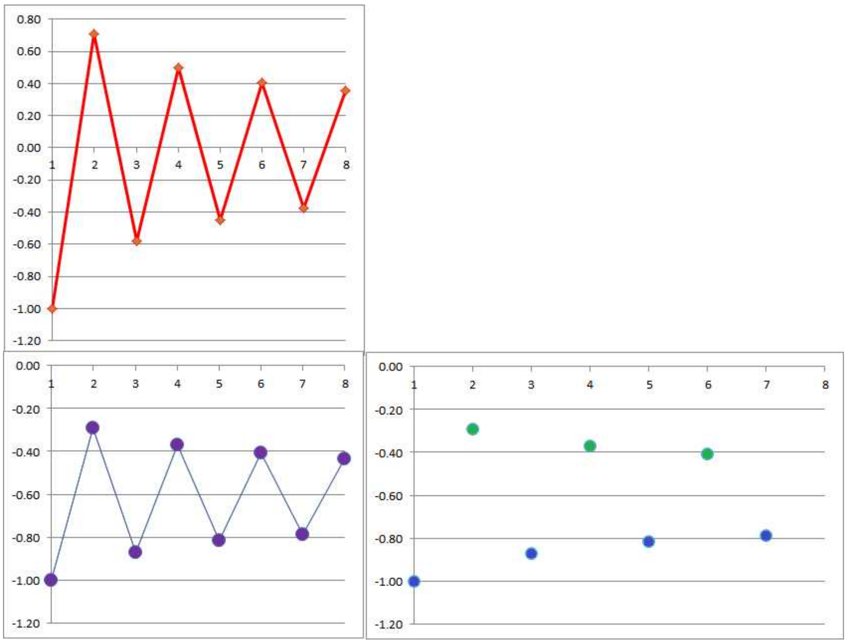
Then the alternating version of the series $\sum b_n$, *the series* $\sum (-1)^n b_n$, *converges.*

Proof.

The *idea* of the proof is as follows. First, the sequence alternates between positive and negative. As a result, the sequence of its sums goes up and down at every step. Furthermore, each step is smaller than the last and the swing is diminishing. Moreover, it is diminishing to zero. That's convergence!

Let's consider the sequence of partial sums of our series:

$$p_n = \sum_{k=1}^{n} (-1)^k b_k \,.$$



We examine the behavior in the subsequences of odd- and even-numbered elements. For the odd:

$$\begin{aligned} p_{2k+1} - p_{2k-1} &= (-1)^{2k} b_{2k} + (-1)^{2k+1} b_{2k+1} \\ &= b_{2k} - b_{2k+1} \\ &> 0 \,. && \text{According to condition 2.} \end{aligned}$$

Therefore,

$$p_{2k+1} \nearrow$$

Similarly, for the even

$$\begin{aligned} p_{2k+2} - p_{2k} &= (-1)^{2k+1} b_{2k+1} + (-1)^{2k+2} b_{2k+2} \\ &= -b_{2k+1} + b_{2k+2} \\ &< 0 \,. && \text{According to condition 2.} \end{aligned}$$

Therefore,

$$p_{2k} \searrow$$

We have two monotone sequences that are also bounded:

$$p_1 \le p_n \le p_2 \,.$$

Therefore, both converge by the *Monotone Convergence Theorem*.

Next, consider these two limits. By the *Squeeze Theorem* we have:

$$\lim_{n\to\infty} (-1)^n b_n = 0 \,,$$

from condition 3. Then, by the *Sum Rule* we have:

$$\lim_{n\to\infty} p_{2k+1} - \lim_{n\to\infty} p_{2k} = \lim_{n\to\infty} (p_{2k+1} - p_{2k}) = \lim_{n\to\infty} (-1)^{2k+1} b_{2k+1} = 0\,.$$

Then, the limits of the odd and the even partial sums are equal.

Therefore, the whole sequence of partial sums p_n converges to the same limit.

Exercise 5.10.12

Provide a proof for the last step.

Corollary 5.10.13

All alternating p-series, i.e.,

$$\sum \frac{(-1)^n}{n^p}\,,$$

converge.

In particular, the alternating p-series,

$$\sum \frac{(-1)^n}{n^p}, \quad \text{with} \quad 0 < p < 1\,,$$

converge *conditionally*.

5.11. The Ratio Test and the Root Test

The challenge of the *Comparison Test* is often how to find a series good for comparison.

In this section, we choose a single type of series and derive all the conclusions we can about the series that compare well with it. This choice is, of course, the *geometric series*.

The most well-understood series is the standard geometric series $\sum r^n$. Its convergence is fully determined by its *ratio* $r \geq 0$:

- If $r < 1$, then $\sum r^n$ converges absolutely.
- If $r > 1$, then $\sum r^n$ diverges.

In other words, the sequence has to go to 0 fast enough for the series to converge.

This idea of the ratio and these two conditions reappear in the case of a generic series.

Indeed, every series $\sum a_n$ has the *ratio*, a sequence:

$$r_n = \frac{a_{n+1}}{a_n}\,.$$

In contrast to the geometric series, the ratio depends on n. But its *limit* does not! As it turns out, the series exhibits the same convergence pattern as the geometric series with the ratio equal to this limit:

Theorem 5.11.1: Ratio Test for Series

Suppose a_n is a sequence with non-zero terms. Suppose the following limit exists, as a number or as infinity:

$$\boxed{r = \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right|}$$

Then we have:

1. *If $r < 1$, then $\sum a_n$ converges absolutely.*
2. *If $r > 1$, then $\sum a_n$ diverges.*

.

If $r = 1$ or the limit doesn't exist, we say that the test fails.

Proof.

Suppose, for a sequence with positive terms, we have:

$$r = \lim_{n\to\infty} \frac{a_{n+1}}{a_n}\,.$$

Then, from the definition of limit, we conclude:

$$\frac{a_{n+1}}{a_n} < s, \text{ for all } n \geq N\,,$$

for some N and *any* $s > r$. Therefore,

$$a_{n+1} < sa_n, \text{ for all } n \geq N\,.$$

This inequality is now applied multiple times, starting at any term $m > N$:

$$a_m < sa_{m-1} < s(sa_{m-2}) = s^2 a_{m-2} < s^2(sa_{m-3}) = s^3 a_{m-3} < ... < s^{m-N} a_N\,.$$

At the end, we have a comparison of our series and a geometric series with ratio s:

$$a_m < \frac{a_N}{s^N} \cdot s^m\,.$$

The latter converges when $s < 1$ and then, by the *Direst Comparison Test*, so does $\sum a_n$. Since s is *any* number above r, this condition is equivalent to: $r < 1$.

Exercise 5.11.2

Complete the proof of the theorem.

Exercise 5.11.3

(a) Give an example of a series the divergence of which the test fails to detect. (b) Give an example of a series the convergence of which the test fails to detect.

Example 5.11.4: exponent and factorial

Let's analyze this series:

$$\sum \frac{2^n}{n!}\,.$$

Its sequence is

$$a_n = \frac{2^n}{n!} .$$

The limit of this sequence is 0; therefore, it passes the *Divergence Test*. This tells us nothing...

Let's turn to the theorem. Our limit is:

$$\begin{aligned} r = \frac{a_{n+1}}{a_n} &= \frac{2^{n+1}/(n+1)!}{2^n/n!} \\ &= \frac{2^{n+1}}{2^n} \cdot \frac{n!}{(n+1)!} \\ &= \frac{2}{1} \cdot \frac{1}{n+1} \\ &\to 0 \qquad\qquad \text{as } n \to \infty. \end{aligned}$$

Therefore, the series converges by the theorem.

What we truly care about, however, are the *power series*! Let's consider one centered at some a:

$$\sum_{n=0}^{\infty} c_n (x-a)^n .$$

Then the theorem applies, for each x, to this numerical sequence:

$$a_n = c_n (x-a)^n ,$$

as follows. We find the limit, for each x:

$$\begin{aligned} r(x) &= \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| \\ &= \lim_{n\to\infty} \left| \frac{c_{n+1}(x-a)^{n+1}}{c_n (x-a)^n} \right| \\ &= \lim_{n\to\infty} \left| \frac{c_{n+1}(x-a)}{c_n} \right| && \text{Factor out by CMR.} \\ &= |x-a| \lim_{n\to\infty} \left| \frac{c_{n+1}}{c_n} \right| \qquad < 1 . && \text{Solve this inequality!} \end{aligned}$$

Then, by the theorem, the series converges for those values of x for which $r(x) < 1$ and diverges for those for which $r(x) > 1$. In the simplest case, we have a positive number:

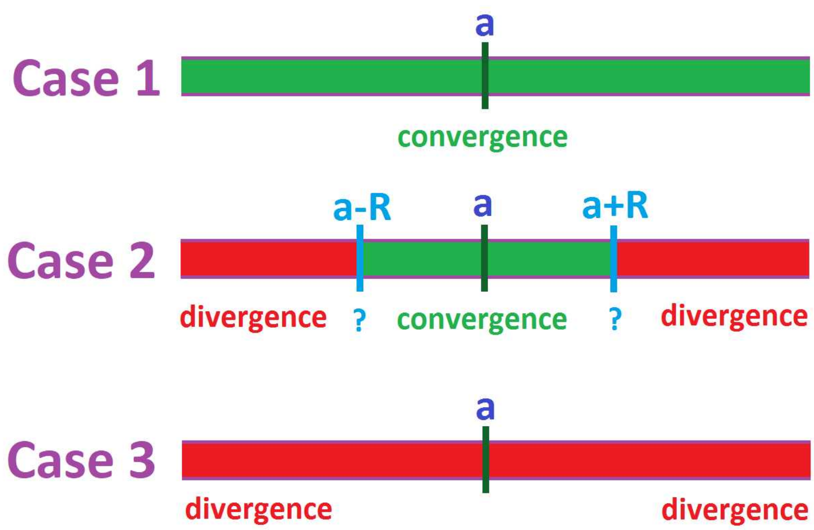
$$R = \frac{1}{\lim_{n\to\infty} \left| \frac{c_{n+1}}{c_n} \right|} .$$

Then, the condition $r(x) < 1$ becomes:

$$|x-a| \frac{1}{R} < 1 \text{ , or } \; |x-a| < R .$$

These x's form an *interval* centered at a:

$$\{x : \; |x-a| < R\} = \{x : \; a - R < x < a + R\} .$$



When our limit is zero, we have $R = \infty$ and when it is infinity, we have $R = 0$.

Corollary 5.11.5: Ratio Test for Power Series

Suppose c_n is a sequence with non-zero terms. Suppose the following limit exists, as a number or as infinity:

$$\boxed{\frac{1}{R} = \lim_{n\to\infty} \left| \frac{c_{n+1}}{c_n} \right|}$$

Then, there are three cases:

	The series $\sum c_n(x-a)^n$ ...
Case 1: $R = \infty$	*converges for each* x;
Case 2: $0 < R < \infty$	*converges for each* x *in the interval:* $(a-R, a+R)$, *and diverges for each* x *in the rays:* $(-\infty, a-R)$, $(a+R, +\infty)$;
Case 3: $R = 0$	*diverges for each* $x \neq a$.

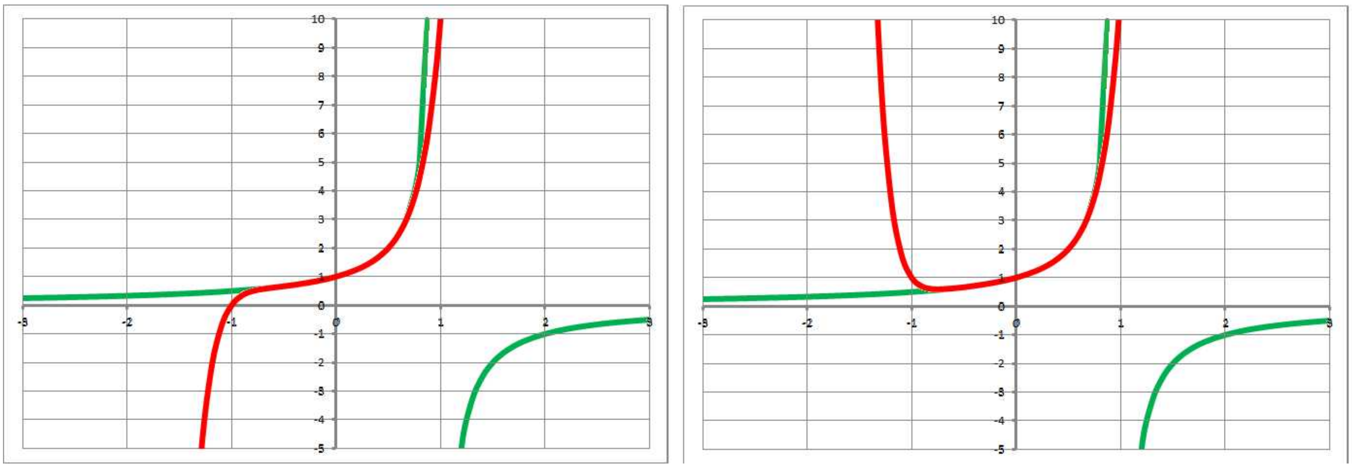
The *end-points* of this interval, $a - R$ and $a + R$, will have to be treated separately.

Example 5.11.6: series of powers

The convergence of $1 + x + x^2 + ...$ is illustrated below. Since $c_n = 1$ and $R = 1$, we infer from the theorem the divergence for x outside the interval $[-1, 1]$ and convergence inside the interval $(-1, 1)$ (the latter is the interior of the former). The two partial sums,

$$1 + x + x^2 + ... + x^9 \quad \text{and} \quad 1 + x + x^2 + ... + x^{10}\,,$$

are shown:



Even though the partial sums are defined for all possible real values of x, the function given by the series:

$$f(x) = 1 + x + x^2 + \dots,$$

is undefined outside the interval $(-1, 1)$.

Example 5.11.7: exponential

Let's confirm the convergence of the Taylor series for $f(x) = e^x$ centered at $a = 0$. We know that this is a power series with

$$c_n = \frac{1}{n!}.$$

Then,

$$c = \lim_{n\to\infty} \left| \frac{c_{n+1}}{c_n} \right| = \lim_{n\to\infty} \frac{1/(n+1)!}{1/n!} = \lim_{n\to\infty} \frac{n!}{(n+1)!} = \lim_{n\to\infty} \frac{1}{n} = 0.$$

The series converges for all x. So, $(-\infty, +\infty)$ is the domain of the function defined by the series.

Next, there is another way to extract the ratio r from a geometric series $a_n = r^n$:

$$r = \sqrt[n]{a_n}.$$

In contrast to the geometric series, this number depends on n. But its *limit* does not! Analogously to the *Ratio Test*, the series exhibits the same convergence pattern as the geometric series with the ratio equal to this limit:

Theorem 5.11.8: Root Test

Suppose a_n is a sequence. Suppose the following limit exists, as a number or as infinity:

$$\boxed{r = \lim_{n\to\infty} \sqrt[n]{|a_n|}}$$

Then, we have:

1. *If $r < 1$, then $\sum a_n$ converges absolutely.*
2. *If $r > 1$, then $\sum a_n$ diverges.*

If $r = 1$ or the limit doesn't exist, the test fails.

Proof.

Suppose, for a sequence with positive terms, we have:

$$r = \lim_{n\to\infty} \sqrt[n]{\frac{a_n}{a_0}}.$$

Then, from the definition of limit, we conclude:

$$\sqrt[n]{\frac{a_n}{a_0}} < R \ \text{ for all } n \geq N\,,$$

for some N and *any* $R > r$. Therefore,

$$a_n < R^n a_0 \ \text{ for all } n \geq N\,.$$

We thus have a comparison of our series and the geometric series with ratio R. The latter converges when $R < 1$ and then, by the *Comparison Test*, so does $\sum a_n$. Since R is *any* number above R, this condition is equivalent to $r < 1$.

Exercise 5.11.9

Complete the proof of the theorem.

Exercise 5.11.10

Examine the convergence of the series $\sum \frac{n}{2^n}$.

Exercise 5.11.11

What is the relation between the two limits in the two theorems?

Example 5.11.12: recursively defined sequences

The *Root Test* requires the nth term sequence to be known! In contrast, the *Ratio Test* can be applied to sequences defined recursively. For example, let

$$a_{n+1} = a_n \cdot \frac{2n+1}{3n+1}\,.$$

There is no direct formula but its convergence is proven by the following computation:

$$\frac{a_{n+1}}{a_n} = \frac{2n+1}{3n+1} \to \frac{2}{3} < 1\,.$$

The series converges.

Let's consider a *power series* again:

$$\sum c_n (x-a)^n\,.$$

Then the theorem applies to this numerical sequence:

$$a_n = c_n (x-a)^n\,,$$

as follows:

$$r(x) = \lim_{n\to\infty} \sqrt[n]{|c_n(x-a)^n|} = |x-a| \lim_{n\to\infty} \sqrt[n]{|c_n|}\,.$$

Then it converges for those values of x for which we have:

$$|x-a| < \frac{1}{\lim_{n\to\infty} \sqrt[n]{|c_n|}}\,.$$

The parameter r is defined via a limit that is quite different from the one in the *Ratio Test* but the conclusions about it are the same.

Exercise 5.11.13

State and prove the Root Test for power series.

5.12. Power series

In the beginning of the chapter, we showed how functions produce power series: via their Taylor polynomials. Since then, we have also demonstrated how power series produce functions: via convergence. We continue below with the latter.

Definition 5.12.1: power series centered at point

A *power series centered at* point $x = a$ is function given by a series that depends on the independent variable x, in the following way:

$$f(x) = \sum_{n=0}^{\infty} c_n (x-a)^n \,,$$

with the sequence of numbers c_n referred to as the *coefficients of the series*.

Every x, the input, once specified, makes the series numerical. If this series converges, we have its sum, a number. It is seen, therefore, as the output of the function.

Just as any function, a power series is a black box:

$$x \ \to \ \boxed{\quad f \quad} \ \to \ y$$

It is a function given by nothing but a *formula*, the formula that might look more complex than the ones we have seen before. This is one of the simplest:

$$f(x) = 1 + x + x^2 + \ldots$$

Just as with any function defined by a formula, the *implied domain* becomes an issue. This is how it works:

- If the series $\sum_{n=0}^{\infty} c_n (x-a)^n$ converges, then x is in the domain of the function.
- If the series $\sum_{n=0}^{\infty} c_n (x-a)^n$ diverges, then x is not in the domain of the function.

Definition 5.12.2: domain of power series

The *domain* of a power series is the set of those x's for which the series converges. It is also called the *region of convergence* of the series.

In other words, we have the following set:

$$\boxed{\text{Domain } = \left\{ x : \sum_{n=0}^{\infty} c_n (x-a)^n \text{ converges} \right\}}$$

Warning!

The value $x = a$ is always in the domain.

Example 5.12.3: geometric

We already know that this familiar power series converges on this interval and diverges outside:

$$1 + x + x^2 + x^3 + ... = \frac{1}{1-x} \quad \text{for all } -1 < x < 1.$$

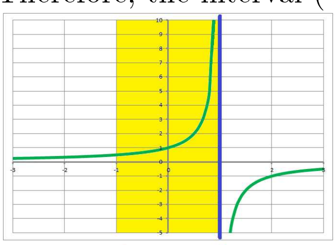
To find the rest of the domain, we consider the end-points of the interval. First:

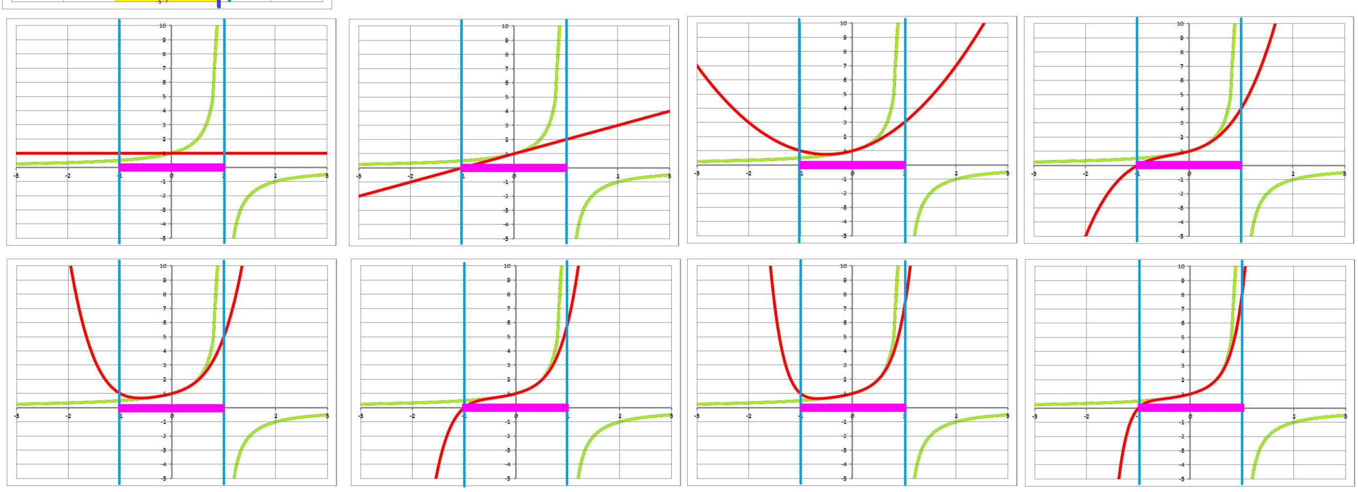
$$1 + x + x^2 + x^3 + ...\Big|_{x=1} = 1 + 1 + 1^2 + 1^3 + ... = 1 + 1 + 1 + 1 + ..., \text{ divergent!}$$

Second:

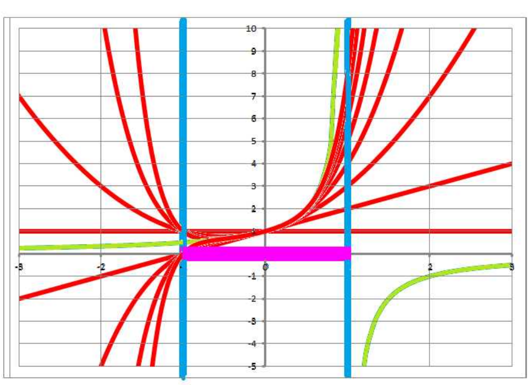
$$1 + x + x^2 + x^3 + ...\Big|_{x=-1} = 1 + (-1) + (-1)^2 + (-1)^3 + ... = 1 - 1 + 1 - 1 + ..., \text{ divergent!}$$

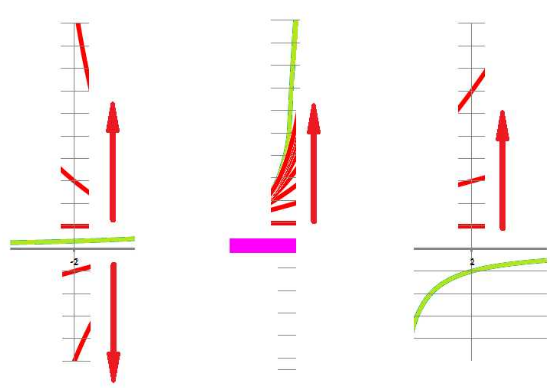
Therefore, the interval $(-1, 1)$ is the whole domain of this function:





We see three cases in the picture below:
1. convergence towards the curve for each x within the interval $(-1, 1)$ (pointwise convergence),
2. divergence that alternates between $+$ and $-$ infinities for each $x < -1$, and
3. divergence away from the curve toward plus infinity for $x > 1$.





Exercise 5.12.4

Prove the statements in the last sentence.

Exercise 5.12.5

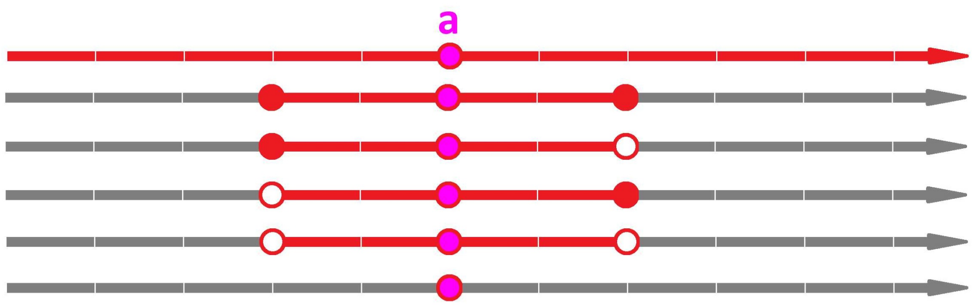
Sketch the divergence at $x = -1$ and $x = 1$.

Warning!

These are two different functions:

$$1 + x + x^2 + x^3 + \ldots \quad \text{and} \quad \frac{1}{1-x}\,.$$

The regions of convergence have been intervals of a special kind:



They are centered on a. The whole line is a possibility. The rest are finite. Those intervals have end-points. Both, one, or neither may belong to the domain. A single point is also a possibility.

We will need the following terminology:

Definition 5.12.6: interior of interval

The *interior* of $\begin{matrix}[a,b]\\ [a,b)\\ (a,b]\\ (a,b)\end{matrix}$ with $a < b$ is (a, b).

The *interior* of $(-\infty, +\infty)$ is $(-\infty, +\infty)$.

Below is a compact summary of the result above:

Theorem 5.12.7: Interval of Convergence

1. *The domain of a power series centered at x is an interval centered at a.*
2. *The convergence is absolute in the interior of the interval of convergence.*
3. *The convergence is uniform on any closed interval in the interior of the interval of convergence.*

Proof.

For simplicity, we assume that $a = 0$. If the only convergent value is $x = 0$, we are done; that's the interval. Suppose now that there is another value, $x = b \neq 0$, i.e.,

$$\sum_{n=0}^{\infty} c_n b^n \text{ converges.}$$

Then, by the *Divergence Test*, we have

$$c_n b^n \to 0\,,$$

and, in particular,

$$|c_n b^n| \leq M \text{ for some } M\,.$$

Therefore, we can estimate our series as follows:

$$|c_n x^n| \leq |c_n b^n| \left|\frac{x}{b}\right|^n \leq M \left|\frac{x}{b}\right|^n\,.$$

The last sequence is a geometric progression and its series converges whenever $|x/b| < 1$, or $|x| < |b|$. Therefore, by the *Direct Comparison Test*, our series converges absolutely for every x in the interval $(-|b|, |b|)$. We conclude that any two points in the domain produce an interval that lies entirely inside the domain. We proved in Chapter 2DC-1 that this property is equivalent to that of being an interval.

In the proof, we can see that the convergence is also absolute.

For the last part, we modify the above proof slightly. Suppose we have a number p with $1 < p < 1$. Then, for any x in the interval $[-p|b|, p|b|]$, we have a comparison of our series with a geometric progression *independent* of x:

$$|c_n x^n| \leq M \left|\frac{x}{b}\right|^n \leq M \frac{|x|}{|b|}^n \leq M \frac{p|b|}{|b|}^n \leq Mp^n\,.$$

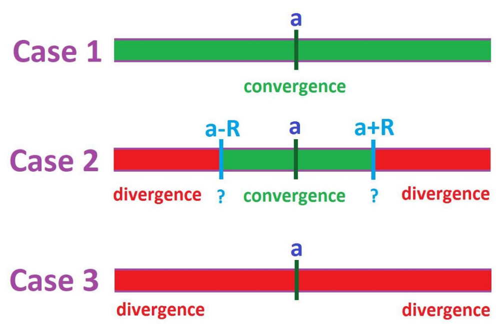
Definition 5.12.8: radius of convergence

The minimal distance R (that could be infinite) from a to the point for which the series diverges is called the *radius of convergence* of the power series centered at a.

This definition is legitimate according to the *Existence of* sup *Theorem* from Chapter 2DC-1.

If R is the radius of convergence, then the length of the domain interval is $2R$ (we will see in Volume 5 that there is, in fact, a *circle* of radius R).

These are the three possibilities:



We make the last theorem more specific below:

Theorem 5.12.9: Radius of Convergence of Power Series

Suppose R is the radius of convergence of a power series:

$$\sum_{n=0}^{\infty} c_n (x-a)^n\,.$$

Then, we have:

1. *When $R = \infty$, the domain of the series is $(-\infty, +\infty)$.*

2. *When $R < \infty$, the domain of the series is an interval with the end-points $a - R$ and $a + R$ (possibly included or excluded).*
3. *When $R = 0$, the domain of the series is $\{a\}$.*

Exercise 5.12.10

Give an example for case 3.

Let's assign domains to power series representations established previously:

series	its sum	its domain
$\sum_{n=0}^{\infty} x^n$	$= \dfrac{1}{1-x}$	$(-1, 1)$
$\sum_{n=0}^{\infty} \dfrac{1}{n!} x^n$	$= e^x$	$(-\infty, +\infty)$
$\sum_{k=0}^{\infty} \dfrac{(-1)^k}{(2k+1)!} x^{2k+1}$	$= \sin x$	$(-\infty, +\infty)$
$\sum_{k=0}^{\infty} \dfrac{(-1)^k}{(2k)!} x^{2k}$	$= \cos x$	$(-\infty, +\infty)$

Exercise 5.12.11

Prove the last two. Hint: Start with the Taylor polynomials.

The main method of finding a power series representation of a function is its *Taylor series*. Other methods are indirect.

Example 5.12.12: power series via substitution

Sometimes we can find the power series representation of a function by ingeniously applying the formula for the *geometric series*, backward:

$$\frac{1}{1-r} = \sum_{n=0}^{\infty} r^n, \text{ for } |r| < 1\,.$$

We have used this idea to find the following representation:

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n\,.$$

Now, the function

$$f(x) = \frac{1}{1-x^2}$$

is recognized as the left-hand side of the first formula with $r = x^2$. Therefore, we can simply write it as a power series (with only even terms):

$$\frac{1}{1-x^2} = \sum_{n=0}^{\infty} (x^2)^n = \sum_{n=0}^{\infty} x^{2n}\,.$$

Similarly, we choose $r = x^3$, after factoring, below:

$$\frac{x}{1-x^3} = x\frac{1}{1-x^3} = x\sum_{n=0}^{\infty}(x^3)^n = \sum_{n=0}^{\infty} x^{3n+1}\,.$$

One more ($r = 2x$):

$$\frac{1}{1-2x} = \sum_{n=0}^{\infty}(2x)^n = \sum_{n=0}^{\infty} 2^n \cdot x^n\,.$$

The center of a series constructed this way may be elsewhere:

$$\frac{1}{x} = \frac{1}{1-(1-x)} = \sum_{n=0}^{\infty}(1-x)^n = \sum_{n=0}^{\infty}(-1)(x-1)^n\,.$$

This method amounts to a *change of variables*.

Exercise 5.12.13

Find the radius of convergence for each of these power series.

The above terminology simplifies the two results established previously.

Theorem 5.12.14: Ratio and Root Tests for Power Series

Suppose c_n is a sequence with non-zero terms. Suppose one of the following two limits exists, as a number or as infinity:

$$R = \lim_{n\to\infty}\left|\frac{c_n}{c_{n+1}}\right| \quad \text{or} \quad R = \frac{1}{\lim_{n\to\infty}\sqrt[n]{|c_n|}}\,.$$

Then, R is the radius of convergence of the power series

$$\sum_{n=0}^{\infty} c_n(x-a)^n\,.$$

Example 5.12.15: radius of convergence

Let's consider

$$\sum \frac{x^n}{n}\,.$$

Following the *Ratio Test*, we need to compute the radius of convergence as the following limit:

$$R = \lim_{n\to\infty}\left|\frac{1/n}{1/(n+1)}\right| = \lim_{n\to\infty}\frac{n+1}{n} = 1\,.$$

Therefore, the interior of the domain is $(-1,1)$. Now the end-points:

$$\begin{array}{llll} x=1 & \Longrightarrow \displaystyle\sum\frac{x^n}{n} & = \displaystyle\sum\frac{1}{n} & \text{is the divergent harmonic series.} \\ x=-1 & \Longrightarrow \displaystyle\sum\frac{x^n}{n} & = \displaystyle\sum\frac{(-1)^n}{n} & \text{is the convergent alternating harmonic series.} \end{array}$$

Therefore, the domain is $[-1,1)$. We have one more point of convergence in comparison to $\sum x^n$.

A given power series produces a function defined on its domain. Because of the *Uniqueness of the Sum of*

Series, it can't produce two!

Let's look at this idea from the other direction:

▶ Can a function have two different representations by a power series?

The issue is handled in the manner that would apply to *polynomials*.

Example 5.12.16: standard form of polynomials

There is the standard way to represent them. For example, $x + x = 2x$ is discovered to be a single representation once we choose to combine the similar terms.

Let's consider a linear function; suppose

$$l(x) = mx + b = nx + c \text{ for each } x\,.$$

Since a linear function can only have one slope and one y-intercept, we conclude that $m = n$ and $b = c$. There is only one representation!

Let's consider a quadratic polynomial; suppose

$$p(x) = ax^2 + bx + c = dx^2 + ex + f \text{ for each } x\,.$$

Do these coefficients have to be the same? Again, we notice that c is the y-intercept, and so is f. They must be equal if this is the same function! We can interpret this geometric observation algebraically. We just plug in $x = 0$ into the equation:

$$p(0) = a0^2 + b0 + c = d0^2 + e0 + f \implies c = f\,.$$

We can now cancel these from the equation for p:

$$p(x) = ax^2 + bx + c = dx^2 + ex + f \implies ax^2 + bx = dx^2 + ex \implies x(ax + b) = x(dx + e)\,.$$

Since it holds for all x, we conclude that we have two equal linear functions:

$$ax + b = dx + e\,.$$

It follows that $a = d$ and $b = e$ according to the above analysis. We can continue on with higher and higher degrees.

The general result is below.

Theorem 5.12.17: Uniqueness of Polynomial Representation

If two polynomials are equal, then their corresponding coefficients in the standard centered form are equal too, i.e.,

$$\begin{array}{lcccl} & \displaystyle\sum_{n=0}^{N} c_n(x-a)^N & = & \displaystyle\sum_{n=0}^{N} d_n(x-a)^n & \text{for all } x \\ \implies & c_n & = & d_n & \text{for all } n = 0, 1, 2, 3, ..., N \end{array}$$

Proof.

The proof is by induction over the degree N. The trick consists of a substitution $x = a$ into this formula; all terms with $(x - a)$ disappear:

$$c_0 = d_0\,.$$

Now these two terms are canceled from our equation, producing:

$$\sum_{n=1}^{N} c_n(x-a)^n = \sum_{n=1}^{N} d_n(x-a)^n .$$

As you can see, the summation starts with $n = 1$ now. The power of $(x - a)$ in every term is then at least 1. We can now factor out $(x - a)$, producing:

$$(x-a)\sum_{n=1}^{N} c_n(x-a)^{n-1} = (x-a)\sum_{n=1}^{N} d_n(x-a)^{n-1} ,$$

or

$$(x-a)\sum_{k=0}^{N-1} c_{k+1}(x-a)^{k} = (x-a)\sum_{k=0}^{N-1} d_{k+1}(x-a)^{k} .$$

Since this holds for all x, the polynomials in parentheses (of degree $N - 1$) are equal. They must have equal coefficients by the inductive assumption.

The idea applies (but not the proof, to be provided in the next section) to power series:

Theorem 5.12.18: Uniqueness of Power Series Representation

If two power series are equal, as functions, on an open interval, then their corresponding coefficients are equal too, i.e.,

$$\begin{array}{rccll} & \displaystyle\sum_{n=0}^{\infty} c_n(x-a)^n & = & \displaystyle\sum_{n=0}^{\infty} d_n(x-a)^n & \text{for all } a-r<x<a+r,\ r>0 \\ \Longrightarrow & c_n & = & d_n & \text{for all } n=0,1,2,3,... \end{array}$$

Warning!

The theorem doesn't follow from the last because the fact that two limits are equal doesn't imply that so are the sequences.

5.13. Calculus of power series

Is there a reason to study power series besides as approximations of functions?

Suppose the center a is given. Then the following items create a function, f:

1. There is a sequence of numbers,
$$c_0, c_1, c_2, ...$$
2. For each input x, its value under this function is computed by substituting it into a formula, a power series with the sequence serving as its coefficients:
$$f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n .$$
3. The domain of the function is the region of convergence of the series.

Power series are functions!

There are may be power series for every occasion, but are they as good as functions?

We start with *algebra*.

Just as with functions in general, we can carry out (some) algebraic operations on power series, producing new power series. However, what is truly important is that we can do these operations *term by term*. The idea comes from our experience with *polynomials*; after all the operations, we want to put the result in the standard form, i.e., with all terms arranged according to the powers.

Example 5.13.1: algebra of polynomials and power series

First, we can add two polynomials one term at a time:

$$\begin{array}{lllll} p(x) & = 1 & +2x & +3x^2 \\ q(x) & = 7 & +5x & -2x^2 \\ \hline p(x)+q(x) & = (1+7) & +(2+5)x & +(3-2)x^2 \end{array}$$

We know that we can also add two series one term at a time (when they converge):

$$\begin{array}{lllll} p(x) & = c_0 & +c_1x & +c_2x^2 & +... \\ q(x) & = d_0 & +d_1x & +d_2x^2 & +... \\ \hline p(x)+q(x) & = (c_0+d_0) & +(c_1+d_1)x & +(c_2+d_2)x^2 & +... \end{array}$$

Second, we can multiply a polynomial by a number one term at a time:

$$\begin{array}{llll} p(x) & = 1 & +2x & +3x^2 \\ \hline 2p(x) & = (2\cdot 1) & +(2\cdot 2)x & +(2\cdot 3)x^2 \end{array}$$

We also multiply a series by a number one term at a time (when it converges):

$$\begin{array}{lllll} p(x) & = c_0 & +c_1x & +c_2x^2 & +... \\ \hline kp(x) & = (kc_0) & +(kc_1)x & +(kc_2)x^2 & +... \end{array}$$

The caveat "when they converge" disappears if the starting point is a function, not a series.

The general result is below:

Theorem 5.13.2: Term-by-Term Algebra of Power Series

Suppose two functions are represented by power series:

$$f(x)=\sum_{n=0}^{\infty}c_n(x-a)^n \quad\text{and}\quad g(x)=\sum_{n=0}^{\infty}d_n(x-a)^n\,.$$

Then we have:

1. *The function $f+g$ is represented by the power series that is the term-by-term sum of those of f and g, defined on the intersection of their domains:*

$$\boxed{(f+g)(x)=\sum_{n=0}^{\infty}c_n(x-a)^n+\sum_{n=0}^{\infty}d_n(x-a)^n=\sum_{n=0}^{\infty}(c_n+d_n)(x-a)^n}$$

2. *The function kf, for any constant k, is represented by the power series that*

is the term-by-term product of that of f, defined on the same domain:

$$(kf)(x) = k \cdot \sum_{n=0}^{\infty} c_n (x-a)^n = \sum_{n=0}^{\infty} (kc_n)(x-a)^n$$

Proof.

The first conclusion is justified by the *Sum Rule for Series.* The second conclusion is justified by the *Constant Multiple Rule for Series.*

In other words, these "infinite" polynomials behave just like ordinary polynomials, wherever they converge.

Next is differentiation and integration, i.e., *calculus*, of power series.

We will see that, just as with functions in general, we can carry out the calculus operations on power series, producing new power series. However, what is truly important is that we can do these operations *term by term.*

Example 5.13.3: calculus of polynomials and power series

Let's differentiate and integrate – only the *Power Formula* required – polynomials.

First, we can differentiate a polynomial one term at a time:

$$\begin{array}{llll} p(x) & = 1 & +2x & +3x^2 \\ \hline p'(x) & = 0 & +2 & +3 \cdot 2x \end{array}$$

Maybe this is just the beginning of a power series:

$$\begin{array}{lllll} p(x) & = 1 & +2x & +3x^2 & +... \\ \hline p'(x) & = 0 & +2 & +3 \cdot 2x & +... \end{array}$$

Second, we can integrate a polynomial one term at a time:

$$\begin{array}{lllll} p(x) & = & 1 & +2x & +3x^2 \\ \hline \displaystyle\int p(x)\, dx & = C & +x & +x^2 & +x^3 \end{array}$$

What if this was just the beginning of a convergent power series?

Example 5.13.4: differentiation

Let's differentiate the terms of the power series representation of the exponential function:

$$\begin{array}{ccccccccccccccc}
e^x & = & 1 & + & x & + & \frac{1}{2!}x^2 & + & \frac{1}{3!}x^3 & + & ... & + & \frac{1}{n!}x^n & + & \frac{1}{(n+1)!}x^{n+1} & +...\\
\downarrow\frac{d}{dx} & & \downarrow\frac{d}{dx} & & \downarrow\frac{d}{dx} & & \downarrow\frac{d}{dx} & & \downarrow\frac{d}{dx} & & & & \downarrow\frac{d}{dx} & & \downarrow\frac{d}{dx} & \\
(e^x)' & \stackrel{?}{=} & 0 & + & 1 & + & \frac{1}{2!}2x & + & \frac{1}{3!}3x^2 & + & ... & + & \frac{1}{n!}nx^{n-1} & + & \frac{1}{(n+1)!}(n+1)x^n & +...\\
 & = & & & 1 & + & x & + & \frac{1}{2!}x^2 & + & ... & + & \frac{1}{(n-1)!}x^{n-1} & + & \frac{1}{n!}x^n & +...\\
 & = & e^x & & & & & & & & & & & & &
\end{array}$$

It works!

Example 5.13.5: integration

Let's integrate the terms of the power series representation of the exponential function:

$$\begin{array}{cccccccccccccc}
e^x & = & & 1 & + & x & + & \frac{1}{2!}x^2 & + & ... & + & \frac{1}{n!}x^n & + & \frac{1}{(n+1)!}x^{n+1} & +...\\
\downarrow\int & & & \downarrow\int & & \downarrow\int & & \downarrow\int & & & & \downarrow\int & & \downarrow\int & \\
\int e^x\,dx & \stackrel{?}{=} & C+ & x & + & \frac{1}{2}x^2 & + & \frac{1}{2!}\frac{1}{3}x^3 & + & ... & + & \frac{1}{n!}\frac{1}{n+1}x^{n+1} & + & \frac{1}{(n+1)!}\frac{1}{n+2}x^{n+2} & +...\\
 & = & C+ & x & + & \frac{1}{2!}x^2 & + & \frac{1}{3!}x^3 & + & ... & + & \frac{1}{(n+1)!}x^{n+1} & + & \frac{1}{(n+2)!}x^{n+2} & +...\\
 & = & K+ & e^x & & & & & & & & & & &
\end{array}$$

It works!

Exercise 5.13.6

Provide details of the above computations.

Differentiation and integration of the terms is easy:

$$\frac{d}{dx}\big(c_n(x-a)^n\big) = nc_n(x-a)^{n-1},\ \int \big(c_n(x-a)^n\big)\,dx = \frac{c_n}{n+1}(x-a)^{n+1}\,.$$

The following theorem is central:

Theorem 5.13.7: Term-by-Term Calculus of Power Series

Suppose the radius of convergence of a power series,

$$f(x) = \sum_{n=0}^{\infty} a_n(x-a)^n\,,$$

is positive or infinite. Then the function f represented by this power series is differentiable (and, therefore, integrable) on the interval of convergence, and the power series representations of its derivative and its antiderivative converge in the interior of this interval and are found by term-by-term differentiation and

integration of the power series of f respectively, i.e.,

$$\boxed{f'(x) = \left(\sum_{n=0}^{\infty} c_n (x-a)^n \right)' = \sum_{n=0}^{\infty} \left(c_n (x-a)^n \right)'}$$

and

$$\boxed{\int f(x)\, dx = \int \left(\sum_{n=0}^{\infty} c_n (x-a)^n \right) dx = \sum_{n=0}^{\infty} \int c_n (x-a)^n \, dx}$$

With this theorem, there is no need for the rules of differentiation or integration except for the *Power Formula*!

Warning!

The theorem speaks of the *interior* of the interval; there is no information about its end-points.

Let's find the explicit formulas for the new series in the standard form:

Corollary 5.13.8: Term-by-Term Differentiation of Power Series

In the interior of the interval of convergence of a power series, we have:

$$f'(x) = \sum_{n=0}^{\infty} \left(c_n (x-a)^n \right)' = \sum_{n=1}^{\infty} n c_n (x-a)^{n-1} = \sum_{k=0}^{\infty} (k+1) c_{k+1} (x-a)^k \,.$$

Note the initial index of 1 instead of 0 in the formula for the derivative.

Corollary 5.13.9: Term-by-Term Integration of Power Series

In the interior of the interval of convergence of a power series, we have:

$$\int f(x)\, dx = \sum_{n=0}^{\infty} \int c_n (x-a)^n \, dx = C + \sum_{n=0}^{\infty} \frac{c_n}{n+1} (x-a)^{n+1} = C + \sum_{k=1}^{\infty} \frac{c_{k-1}}{k} (x-a)^k \,.$$

Example 5.13.10: differential equations

In contrast to the last example, let's use the theorem to "discover" a solution rather than confirm it.

Suppose we have a model that states that a quantity's rate of change is proportional (or equal at its simplest) to the current values of the quantity. We need to solve this differential equation:

$$f' = f \,.$$

We assume that the unknown function $y = f(x)$ is represented by a power series. We differentiate the

series and then match the terms according to the equation:

$$\begin{array}{cccccccccccccc}
f & = & c_0 & + & c_1x & + & c_2x^2 & + & c_3x^3 & + & ... & + & c_nx^n & +...\\
f' & = & & & c_1 & + & 2c_2x & + & 3c_3x^2 & + & ... & + & nc_nx^{n-1} & +...\\
\Longrightarrow & & & \swarrow & & \swarrow & & \swarrow & & ... & & \swarrow & & \\
f' & = & c_1 & + & 2c_2x & + & 3c_3x^2 & + & ... & + & nc_nx^{n-1} & + & (n+1)c_{n+1}x^n & +...\\
|| & & || & & || & & || & & & & || & & || & \\
f & = & c_0 & + & c_1x & + & c_2x^2 & + & ... & + & c_{n-1}x^{n-1} & + & c_nx^n & +...
\end{array}$$

According to the *Uniqueness of Power Series*, the coefficients have to be equal!

We have created a sequence of equations:

$$\begin{array}{cccccc}
c_1 & 2c_2 & 3c_3 & ... & (n+1)c_{n+1} & ...\\
|| & || & || & & || & \\
c_0 & c_1 & c_2 & ... & c_n & ...
\end{array}$$

We have converted a calculus problem into an algebra problem!

We face infinitely many equations with infinitely many unknowns. We start solving one equation at a time and substitute the result into the next equation, moving from left to right:

$$\begin{array}{lllllllllll}
c_1 & \Longrightarrow & c_1 = c_0 & 2c_2 & \Longrightarrow & c_2 = c_0/2 & 3c_3 & \Longrightarrow & c_3 = c_0/(2\cdot 3) & 4c_4 & \\
|| & & & || & & & || & & & || & \\
c_0 & & \Longrightarrow & c_1 = c_0 & & \Longrightarrow & c_2 = c_0/2 & & \Longrightarrow & c_3 = ... &
\end{array}$$

The pattern becomes clear:

$$c_{n+1} = \frac{c_n}{n}\,.$$

Therefore,

$$c_n = \frac{c_0}{n!}\,.$$

The problem is solved! Indeed, we have a formula for the function we are looking for:

$$f(x) = \sum_{n=0}^{\infty} \frac{c_0}{n!}x^n = c_0 \sum_{n=0}^{\infty} \frac{1}{n!}x^n\,,$$

and it will give us the values of f with any accuracy we want. The only missing part in this program is the proof of *convergence*; it is done with the *Ratio Test* (seen previously): $R = \infty$. As a *bonus* (just a bonus!), we recognize the resulting series: $f(x) = c_0e^x$.

Exercise 5.13.11

Apply the method to the differential equations: (a) $f' = 2f$; (b) $f' = f + 1$; (c) $f'' = f$; (d) $f'' = -f$.

Exercise 5.13.12

Show that the method doesn't work for $f' = f^2$.

This method of solving differential equations is further discussed in Chapter 5DE-2.

The new approach to calculus is based on the following concept:

Definition 5.13.13: analytic function

A function defined on an open interval that can be represented by a power series is called *analytic* on this interval.

We can finally prove the important result we put forward in the last section:

Corollary 5.13.14: Uniqueness of Power Series Representation

An analytic function has a unique power series representation, i.e., if two power series are equal, as functions, on an open interval, then their corresponding coefficients are equal too, i.e.,

$$\begin{array}{lrcll} f(x) & =\sum_{n=0}^{\infty} c_n(x-a)^n & = & \sum_{n=0}^{\infty} d_n(x-a)^n & \text{for all } a-r<x<a+r,\ r>0 \\ \Longrightarrow & c_n & = & d_n & \text{for all } n=0,1,2,3,... \end{array}$$

Proof.

It's the same function; therefore, the value at $x=a$ is the same:

$$\begin{array}{lll} f(a) & =\sum_{n=0}^{\infty} c_n(a-a)^n & = \sum_{n=0}^{\infty} d_n(a-a)^n \Longrightarrow \\ f(a) & =c_0 & = d_0 \end{array}$$

It's the same function; therefore, the value of the *derivative* at $x=a$ is the same:

$$\begin{array}{lll} f'(a) & =\sum_{n=1}^{\infty} c_n n(a-a)^{n-1} & = \sum_{n=1}^{\infty} d_n n(a-a)^{n-1} \Longrightarrow \\ f'(a) & =c_1 & = d_1 \end{array}$$

Then the second derivative, so on.

Corollary 5.13.15

Every analytic function is infinitely many times differentiable.

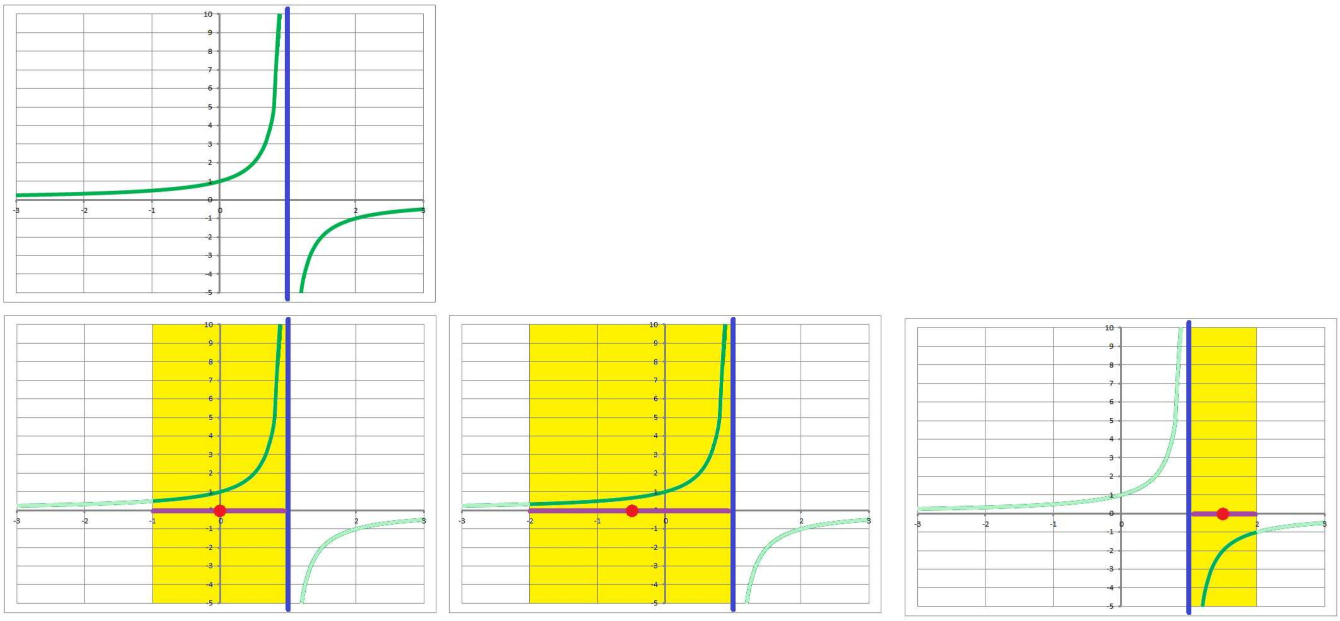
What about the converse? What kind of functions can be represented by power series?

Example 5.13.16: domain of power series

Consider again:

$$f(x)=\frac{1}{x-1}\,.$$

The picture illustrates the limitations of representing functions by power series. The centers we try are $a=0,\ -.5,\ .5$:



Large parts of the graphs lie outside the strip every time. Why? Because the domain of any of these power series is an interval centered at a that cannot contain 1! The good news is that together they cover the whole graph.

The term-by-term differentiation allows us to rediscover the *Taylor polynomials*. Suppose we already have a power series representation of a function with a radius of convergence $R > 0$:

$$f(x) = c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3 + ... + c_n(x-a)^n + ...$$

Let's express the coefficients in terms of the function itself. The trick is partly the same as in the proof of the last theorem: We substitute $x = a$ into this and the derived formulas, except this time we don't divide but rather differentiate. First substitution gives us what we already know:

$$f(a) = c_0\,.$$

We have found c_0. We differentiate,

$$f'(x) = c_1 + 2c_2(x-a) + 3c_3(x-a)^2 + ... + nc_n(x-a)^{n-1} + ...\,,$$

and substitute $x = a$, giving us:

$$f'(a) = c_1\,.$$

We have found c_1. We differentiate one more time,

$$f''(x) = 2c_2 + 3c_3(x-a) + ... + n(n-1)c_n(x-a)^{n-2} + ...\,,$$

and substitute $x = a$, giving us:

$$f''(a) = 2c_2\,.$$

After m steps, we have:

$$f^{(m)}(x) = m(m-1)...3\cdot 2\cdot 1\cdot c_m + m(m-1)...3\cdot 2\cdot c_{m-1}(x-a) + \ ...\ + n\cdot(n-1)\cdots(n-m)\cdot c_n(x-a)^{n-m} + \ ...\,,$$

and substituting $x = a$ gives us:

$$f^{(m)}(a) = m!c_m\,.$$

We have found c_m.

The result of this computation is presented below:

Theorem 5.13.17: Taylor Coefficients

If a function is represented by a power series with a positive radius of convergence,

$$f(x) = c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3 + \ldots + c_n(x-a)^n + \ldots,$$

then its coefficients are the Taylor coefficients:

$$c_n = \frac{f^{(n)}(a)}{n!}.$$

Thus, the nth partial sum of this series is the nth Taylor polynomial of f.

Definition 5.13.18: Taylor series

Suppose f is an infinitely differentiable function. Then the power series

$$c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3 + \ldots + c_n(x-a)^n + \ldots,$$

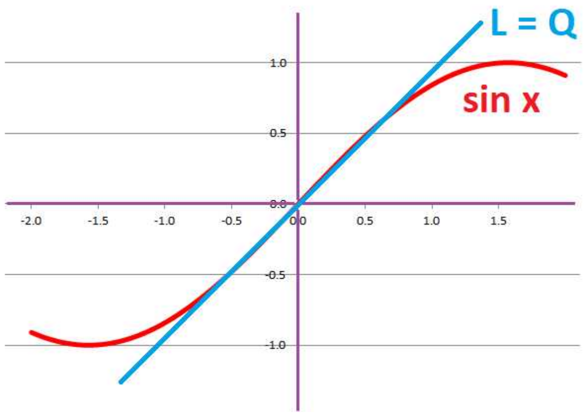
with coefficients:

$$c_n = \frac{f^{(n)}(a)}{n!}$$

is called the *Taylor series* of f at $x = a$.

Example 5.13.19: sine and cosine

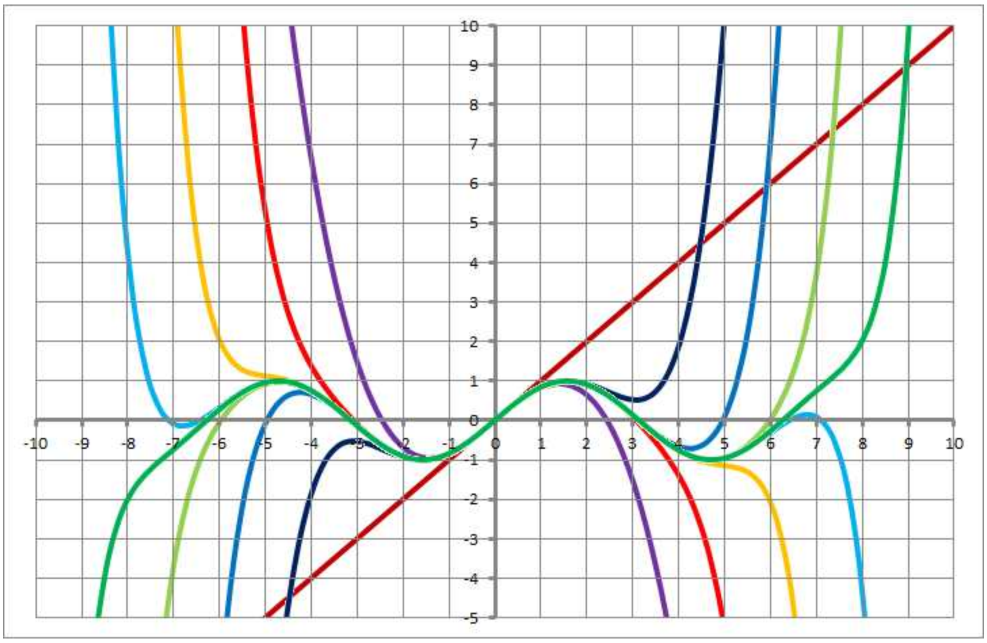
Let's find the power series representation for $f(x) = \sin x$ at $x = 0$. We start with what we already know:



We need them all:

$$\begin{array}{llllllll}
f(x) & = \sin x & \Longrightarrow & f(0) & = 0 & \Longrightarrow & T_0(x) & = 0 \\
f'(x) & = \cos x & \Longrightarrow & f'(0) & = 1 & \Longrightarrow & T_1(x) & = x \\
f''(x) & = -\sin x & \Longrightarrow & f''(0) & = 0 & \Longrightarrow & T_2(x) & = x \\
f'''(x) & = -\cos x & \Longrightarrow & f'''(0) & = -1 & \Longrightarrow & T_3(x) & = 1 - \dfrac{1}{6}x^3 \\
\ldots & & & & & & &
\end{array}$$

The sequence starts to repeat itself, every four steps. Of course, every polynomial leaves for infinity eventually, but the resemblance extends further and further from the center:



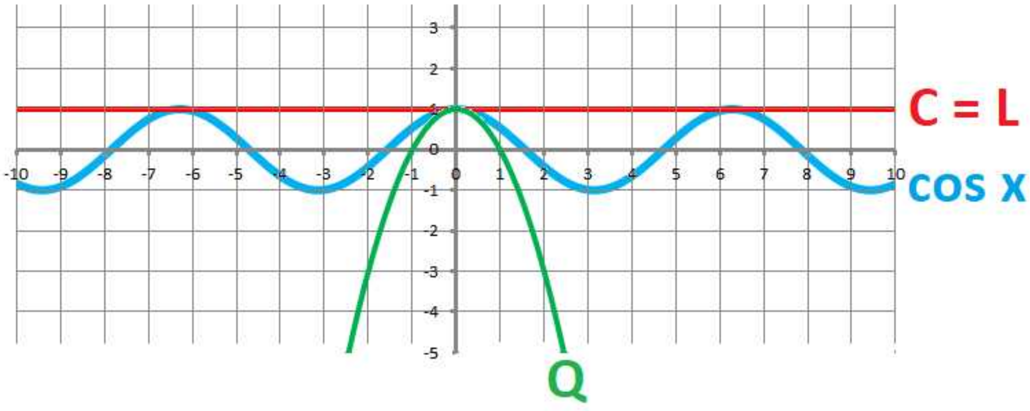
There are no even powers present because the sine is odd. Therefore,

$$f^{(2m-1)}(0) = (-1)^m \,.$$

We have the Taylor coefficients:

$$c_{2m-1} = \frac{(-1)^m}{(2m-1)!}\,.$$

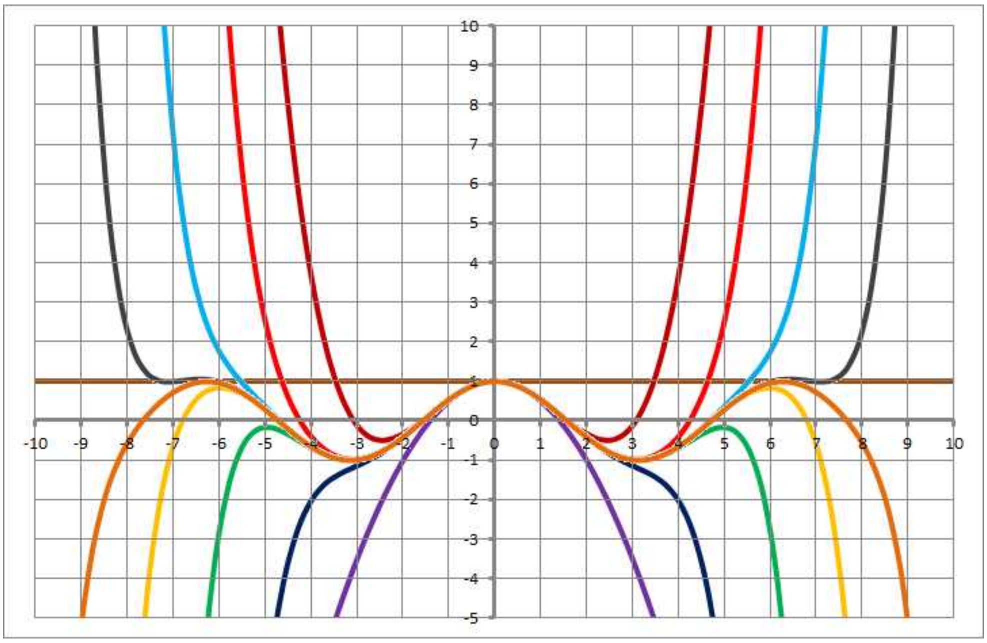
Let's approximate $f(x) = \cos x$ at $x = 0$. We start with what we already know:



We need them all:

$$\begin{array}{llllll}
f(x) & = \cos x & \Longrightarrow & f(0) & = 1 & \Longrightarrow & T_0(x) & = 1 \\
f'(x) & = -\sin x & \Longrightarrow & f'(0) & = 0 & \Longrightarrow & T_1(x) & = 1 \\
f''(x) & = -\cos x & \Longrightarrow & f''(0) & = -1 & \Longrightarrow & T_2(x) & = 1 - \dfrac{1}{2}x^2 \\
f'''(x) & = \sin x & \Longrightarrow & f'''(0) & = 0 & \Longrightarrow & T_3(x) & = 1 - \dfrac{1}{2}x^2 \\
f^{(4)}(x) & = \cos x & \Longrightarrow & f^{(4)}(0) & = 1 & & & \\
\ldots & & & & & & &
\end{array}$$

The sequence starts to repeat itself, every four steps. Once again, every polynomial leaves for infinity eventually, but the resemblance extends further and further from the center:



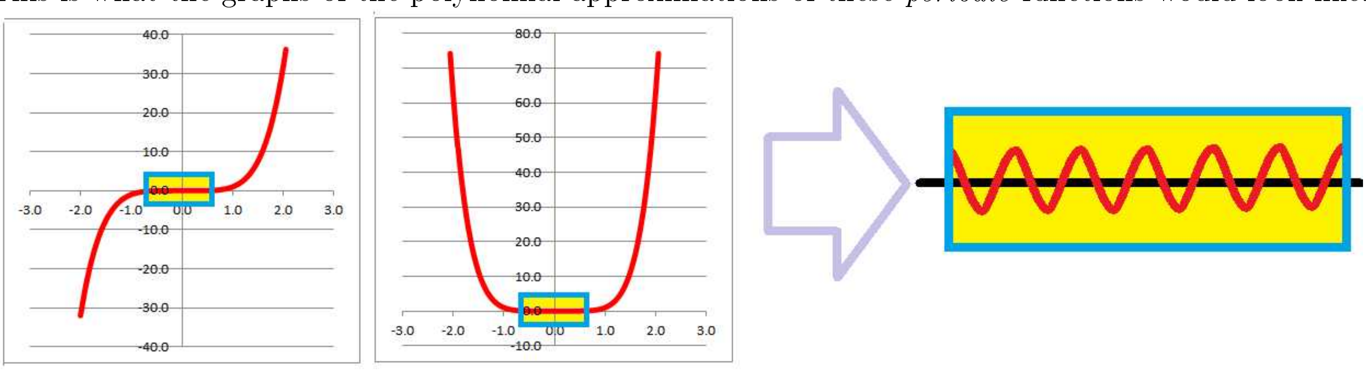
There are no odd powers present because the cosine is even. Therefore,

$$f^{(2m)}(0) = (-1)^m .$$

We have the Taylor coefficients:

$$c_{2m} = \frac{(-1)^m}{(2m)!} .$$

This is what the graphs of the polynomial approximations of these *periodic* functions would look like:



They eventually become bad, but the interval where things are good is expanding.

The surprising byproduct of the theorem is the conclusion that the whole analytic function is determined by the values of its derivatives *at a single point*.

Corollary 5.13.20

Suppose f, g are analytic on interval $(a-R, a+R)$, $R > 0$. If they have matching derivatives of all orders at a, they are equal on this interval; i.e.,

$$f^{(n)}(a) = g^{(n)}(a) \text{ for all } n = 0, 1, 2, ... \Longrightarrow f(x) = g(x) \text{ for all } x \text{ in } (a-R, a+R) .$$

Exercise 5.13.21

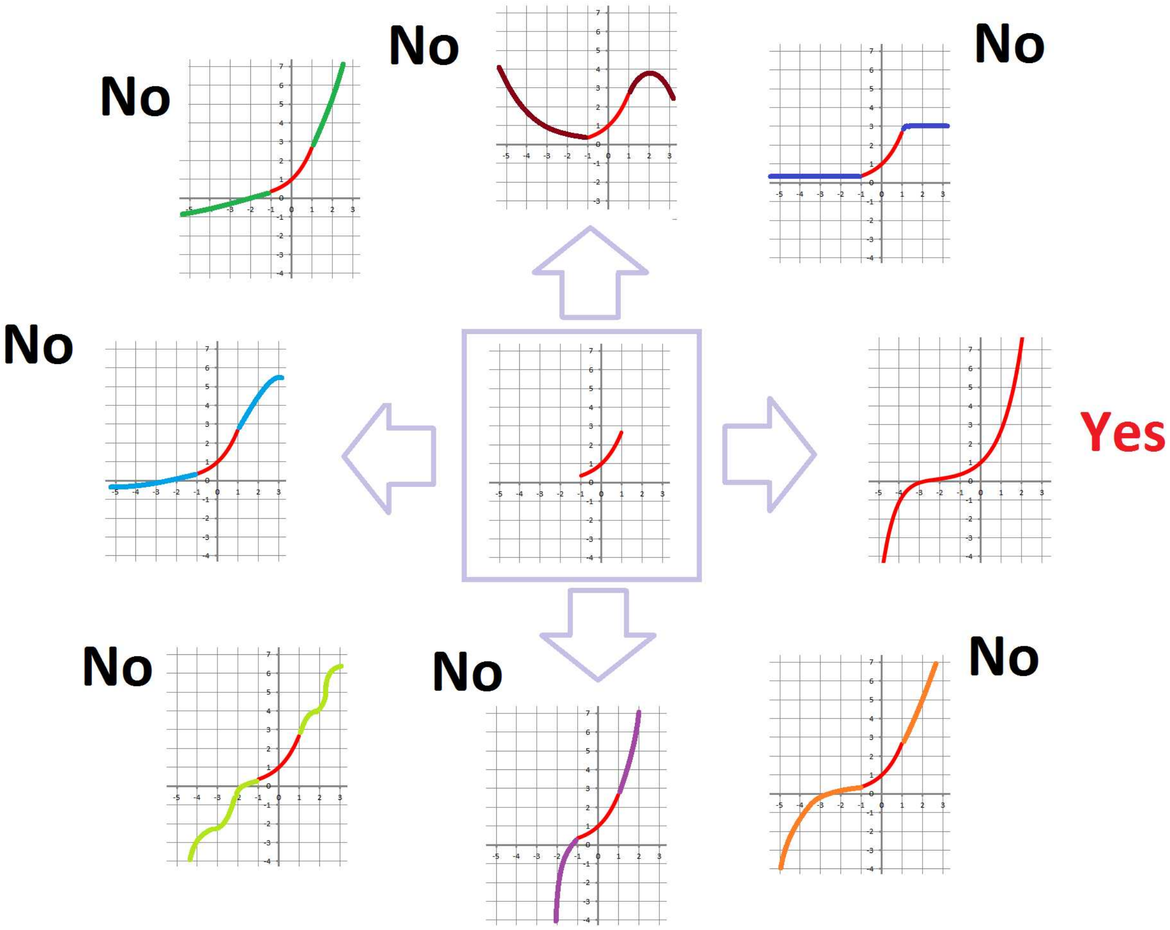
Prove the corollary. Hint: Consider $f - g$.

Since these derivatives are, in turn, fully determined by the behavior of the function on a small (no matter how small) interval around this point, we conclude:

► There is only *one way* to extend the function beyond this interval.

For example, a function constant around the point will have to be constant everywhere else. A function linear around a point will have to be linear everywhere else. And the same for the quadratic functions too.

Below is the general situation:



In other words, analytic functions are extremely "predictable": Once we have drawn a tiny piece of the graph, there is only one way to continue to draw it. We can informally interpret this idea as follows:

▶ Drawing a curve with a single stroke of the pen produces an analytic function, but stopping in the middle to decide how to proceed is likely to prevent this from happening.

What kind of functions can be represented by power series?

Theorem 5.13.22: Representation by Power Series

Suppose a function f is infinitely many times differentiable on interval $(a-R, a+R)$, and these derivatives are bounded by the same constant M:

$$|f^{(n)}(x)| \le M \quad \text{for all } x \text{ in } (a-R, a+R)\,.$$

Then f is analytic.

Proof.

Just compare the Taylor series of the function with this convergent series:

$$\sum \frac{M}{n!}\,.$$

Exercise 5.13.23

Weaken the boundedness condition.

Example 5.13.24: Taylor series via substitutions

Find the power series representation of

$$f(x)=x^{-3}$$

around $a=1$. We differentiate and substitute:

$$\begin{array}{lllll}
f(x) & =x^{-3} & \Rightarrow\ f(1)=1 & \Rightarrow\ c_0=1 \\
f'(x) & =(-3)x^{-4} & \Rightarrow\ f'(1)=-3 & \Rightarrow\ c_1=-3 \\
f''(x) & =(-3)(-4)x^{-5} & \Rightarrow\ f''(1)=12 & \Rightarrow\ c_2=12/2=6 \\
f^{(3)}(x) & =(-3)(-4)(-5)x^{-6} & \Rightarrow\ f^{(3)}(1)=-60 & \Rightarrow\ c_3=-60/6=-10 \\
.. & ... & ...\ ... & ...\ ... \\
f^{(n)}(x) & =(-3)(-4)\cdots(-2-n)x^{-3-n} & \Rightarrow\ f^{(n)}(1)=\dfrac{(-1)^n(2+n)!}{2} & \Rightarrow\ c_n=(-1)^n\dfrac{(n+1)(n+2)}{2}
\end{array}$$

That's the nth Taylor coefficient. Hence,

$$f(x)=\sum_{n=1}^{\infty}(-1)^n\frac{(n+1)(n+2)}{2}(x-a)^n\,.$$

But what is the interval of convergence? The radius of convergence is the distance to the nearest point for which there is no convergence, which appears to be 0, so $R=1$. We confirm this with the *Ratio Test*:

$$\lim_{n\to\infty}\left|(-1)^n\frac{(n+1)(n+2)}{2}\div(-1)^{n+1}\frac{(n+2)(n+3)}{2}\right|=\lim_{n\to\infty}\frac{(n+1)(n+2)}{(n+2)(n+3)}=1\,.$$

Then the end-points of the interval are 0 and 2. The series converges in the interior of the interval, and what's left is the convergence at the end-points:

$$\begin{array}{lll}
x=0, & \sum(n+1)(n+2) & \text{diverges.}\\
x=2, & \sum(n+1)(n+2)(-1)^n & \text{diverges.}
\end{array}$$

Thus, $(0,2)$ is the interval of convergence of the series. It is also the domain of the series even though it's smaller than that of the original function.

The mismatch between the domains of functions and the intervals of convergence of their power series representation is the (main) reason why the match between calculus of functions and calculus of power series is imperfect.

At least, the differentiation and integration, according to the *Term-by-Term Differentiation and Integration*, are perfectly reflected in this mirror:

$$\begin{tikzcd}
f \arrow[r, "\text{Taylor}"] \arrow[d, "\frac{d}{dx}"] & \sum c_n(x-a)^n \arrow[d, "\frac{d}{dx}"] \\
f' \arrow[r, "\text{Taylor}"] & \sum\left(c_n(x-a)^n\right)'
\end{tikzcd}
\qquad
\begin{tikzcd}
f \arrow[r, "\text{Taylor}"] \arrow[d, "\int"] & \sum c_n(x-a)^n \arrow[d, "\int"] \\
\int f\,dx \arrow[r, "\text{Taylor}"] & \sum\int\left(c_n(x-a)^n\right)\,dx
\end{tikzcd}$$

In the first diagram, we start with a function at the top left and then we proceed in two ways:

- Right: find its Taylor series. Then down: differentiate the result term by term.
- Down: differentiate it. Then right: find its Taylor series.

The result is the same!

This study will continue with further applications in Chapter 5DE-2.

Exercises

Contents

1. Exercises: Background

Exercise 1.1

What are the max, min, and any bounds of the set of integers? What about $\mathbf{R}$?

Exercise 1.2

Is the converse of a true statement true?

Exercise 1.3

Is the converse of the converse of a true statement true?

Exercise 1.4

State the converse of this statement: "The converse of the converse of a true statement is true".

Exercise 1.5

Represent these sets as intersections and unions:

1. $(0,5)$
2. $\{3\}$
3. $\varnothing$
4. $\{x : x > 0 \ \texttt{ OR } \ x \text{ is an integer}\}$

5. $\{x : x \text{ is divisible by } 6\}$

Exercise 1.6

True or false: $0 = 1 \implies 0 = 1$?

Exercise 1.7

Prove or disprove:

$$\max\{\max A, \max B\} = \max(A \cup B)\,.$$

Exercise 1.8

(a) If, starting with a statement A, after a series of conclusions you arrive to $0 = 1$, what can you conclude about A? (b) If, starting with a statement A, after a series of conclusions you arrive to $0 = 0$, what can you conclude about A?

Exercise 1.9

We know that "If it rains, the road gets wet". Does it mean that if the road is wet, it has rained?

Exercise 1.10

A garage light is controlled by a switch and, also, it may automatically turn on when it senses motion during nighttime. If the light is OFF, what do you conclude?

Exercise 1.11

If an advertisement claims that "All our second-hand cars come with working AC", what is the easiest way to disprove the sentence?

Exercise 1.12

Teachers often say to the student's parents: "If your student works harder, he'll improve". When he won't improve and the parents come back to the teacher, he will answer: "He didn't improve, that means he didn't work harder". Analyze.

Exercise 1.13

Suppose the cost is $f(x)$ dollars for a taxi trip of x miles. Interpret the following stories in terms of f.

1. Monday, I took a taxi to the station 5 miles away.
2. Tuesday, I took a taxi to the station but then realized that I left something at home and had to come back.
3. Wednesday, I took a taxi to the station and I gave my driver a five dollar tip.
4. Thursday, I took a taxi to the station but the driver got lost and drove five extra miles.
5. Friday, I have been taking a taxi to the station all week on credit; I pay what I owe today.

What if there is an extra charge per ride of m dollars?

Exercise 1.14

Let $f : A \to B$ and $g : C \to D$ be two possible functions. For each of the following functions, state whether or not you can compute $f \circ g$:

- $D \subset B$
- $C \subset A$
- $B \subset D$
- $B = C$

Exercise 1.15

Function $y = f(x)$ is given below by a list of some of its values. Make sure the function is onto.

x	-1	0	1	2	3	4	5
$y = f(x)$	-1		4	5		2	

Exercise 1.16

Function $y = f(x)$ is given below by a list of some of its values. Add missing values in such a way that the function is one-to-one.

x	-1	0	1	2	3	4	5
$y = f(x)$	-1		0	5		0	

Exercise 1.17

Represent the function $h(x) = \sin^2 x + \sin^3 x$ as the composition $g \circ f$ of two functions $y = f(x)$ and $z = g(y)$.

Exercise 1.18

Function $y = f(x)$ is given below by a list its values. Find its inverse and represent it by a similar table.

x	0	1	2	3	4
$y = f(x)$	0	1	2	4	3

Exercise 1.19

Find the formulas of the inverses of the following functions: (a) $f(x) = (x + 1)^3$; (b) $g(x) = \ln(x^3)$.

Exercise 1.20

Given the tables of values of f, g, find the table of values of $f \circ g$:

x	$y = g(x)$
0	0
1	4
2	3
3	0
4	1

y	$z = f(y)$
0	4
1	4
2	0
3	1
4	2

What if the last rows were missing?

Exercise 1.21

Represent the function below as the composition $f \circ g$ of two functions:

$$h(x) = \sqrt{2x^3 + x}\,.$$

Exercise 1.22

Represent the function $h(x) = 2\sin^3 x + \sin x + 5$ as the composition of two functions one of which is trigonometric.

Exercise 1.23

Suppose a function f performs the operation: "take the logarithm base 2 of", and function g performs: "take the square root of". (a) Verbally describe the inverses of f and g. (b) Find the formulas for these four functions. (c) Give them domains and codomains.

Exercise 1.24

1. Represent the function $h(x) = \sqrt{x^2 - 1}$ as the composition of two functions f and g.
2. Provide a formula for the composition $y = f(g(x))$ of $f(u) = u^2 + u$ and $g(x) = 2x - 1$.

Exercise 1.25

Provide a formula for the composition $y = f(g(x))$ of $f(u) = \sin u$ and $g(x) = \sqrt{x}$.

Exercise 1.26

Provide a formula for the composition $y = f(g(x))$ of $f(u) = u^2 - 3u + 2$ and $g(x) = x$.

Exercise 1.27

What is the meaning of the inverse of the function $f(x) = 3x^2 + 1$? Hint: Choose appropriate domains.

Exercise 1.28

(a) What is the composition $f \circ g$ for the functions given by $f(u) = u^2 + u$ and $g(x) = 3$? (a) What is the composition $f \circ g$ for the functions given by $f(u) = 2$ and $g(x) = \sqrt{x}$?

Exercise 1.29

Function $y = f(x)$ is given below by a list its values. Find its inverse and represent it by a similar table.

x	0	1	2	3	4
$y = f(x)$	1	2	0	4	3

Exercise 1.30

Give examples of functions that are their own inverses.

Exercise 1.31

Represent this function: $h(x) = \tan(2x)$ as the composition of two functions of variables x and y.

Exercise 1.32

Represent this function:

$$h(x) = \frac{x^3 + 1}{x^3 - 1}\,,$$

as the composition of two functions of variables x and y.

Exercise 1.33

Function $y = f(x)$ is given below by a list of its values. Is the function one-to one? What about its inverse?

x	0	1	2	3	4
$y = f(x)$	0	1	2	1	2

Exercise 1.34

Function $y = f(x)$ is given below by a list of its values. Is the function one-to one? What about its inverse?

x	0	1	2	3	4
$y = f(x)$	7	5	3	4	6

Exercise 1.35

Functions $y = f(x)$ and $u = g(y)$ are given below by tables of some of their values. Present the composition $u = h(x)$ of these functions by a similar

table:

x	0	1	2	3	4
$y = f(x)$	1	1	2	0	2

y	0	1	2	3	4
$u = g(y)$	3	1	2	1	0

Exercise 1.36

(a) Algebraically, show that the function $f(x) = x^2$ is not one-to-one. (b) Graphically, show that the function $g(x) = 2^{x+1}$ is one-to-one. (c) Find the inverse of g.

Exercise 1.37

Describe – both geometrically and algebraically – two different transformations that make a 1×1 square into a 2×3 rectangle.

Exercise 1.38

Find the equation of the line passing through the points $(-1, 1)$ and $(-1, 5)$.

Exercise 1.39

Consider triangle ABC in the plane where $A = (3, 2)$, $B = (3, -3)$, $C = (-2, -2)$. Find the lengths of the sides of the triangle.

Exercise 1.40

Find all x such that the distance between the points $(3, -8)$ and $(x, -6)$ is 5.

Exercise 1.41

Find the distance between the points of intersection of the circle $(x - 1)^2 + (y - 2)^2 = 6$ with the axes.

Exercise 1.42

Solve the system of linear equations:

$$\begin{cases} x - y = -1, \\ 2x + y = 0. \end{cases}$$

Exercise 1.43

A movie theater charges \$10 for adults and \$6 for children. On a particular day when 320 people paid an admission, the total receipts were \$3120. How many were adults and how many were children?

Exercise 1.44

The taxi charges \$1.75 for the first quarter of a mile and \$0.35 for each additional fifth of a mile. Find a linear function which models the taxi fare f as a function of the number of miles driven, x.

Exercise 1.45

Find the value of k so that the line containing the points $(-6, 0)$ and $(k, -5)$ is parallel to the line containing the points $(4, 3)$ and $(1, 7)$.

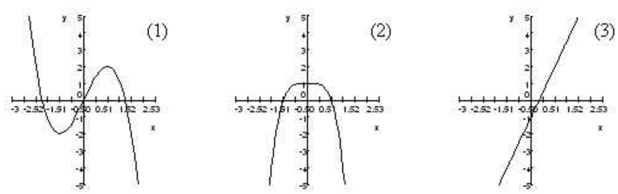
Exercise 1.46

Find the equation satisfied by all points that lie 2 units away from the point $(-1, -2)$ and by no other points.

Exercise 1.47

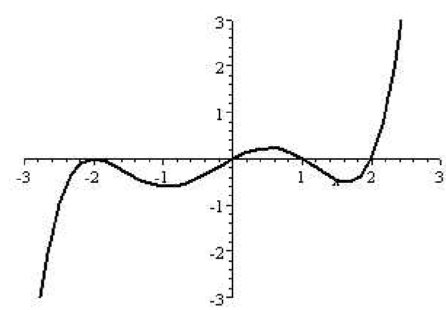
For the polynomials graphed below, find the following:

	1	2	3
smallest possible degree			
sign of the leading coefficient			
degree is odd/even			



Exercise 1.48

Find a possible formula for the function plotted below:



Exercise 1.49

A factory is to be built on a lot measuring 240 ft by 320 ft. A building code requires that a lawn of uniform width and equal in area to the factory must surround the factory. What must the width of the lawn be?

Exercise 1.50

A factory occupies a lot measuring 240 ft by 320 ft. A building code requires that a lawn of uniform width and equal in area to the factory must surround the factory. What must the width of the lawn be?

Exercise 1.51

Make a flowchart and then provide a formula for the function $y = f(x)$ that represents a parking fee for a stay of x hours. It is computed as follows: free for the first hour and \$1 per hour beyond.

Exercise 1.52

Explain the difference between these two functions:

$$\sqrt{\frac{x-1}{x+1}} \quad \text{and} \quad \frac{\sqrt{x-1}}{\sqrt{x+1}}\,.$$

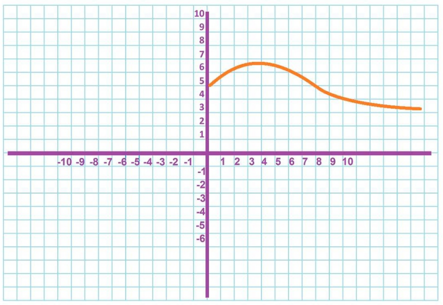
Exercise 1.53

Classify these functions:

function	odd	even	onto	one-to-one
$f(x) = 2x - 1$				
$g(x) = -x + 2$				
$h(x) = 3$				

Exercise 1.54

The graph of $y = f(x)$ is plotted below. Sketch $y = -f(x+5) - 6$.



Exercise 1.55

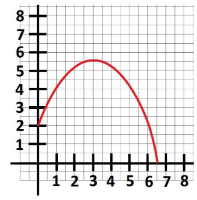
Is the composition of two functions that are odd-/even odd/even?

Exercise 1.56

Find the formulas of the inverses of the following functions: (a) $f(x) = (x+1)^3$; (b) $g(x) = \ln(x^3)$.

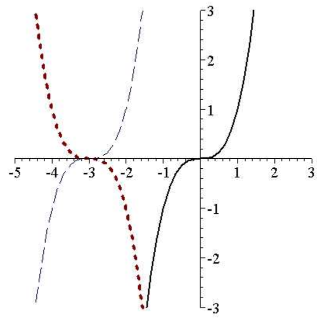
Exercise 1.57

The graph of the function $y = f(x)$ is given below. Sketch the graph of $y = 2f(3x)$ and then $y = f(-x) - 1$.



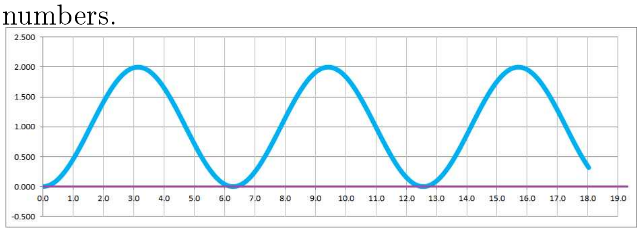
Exercise 1.58

The graph drawn with a solid line is $y = x^3$. What are the other two?



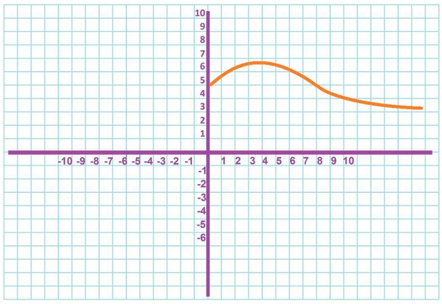
Exercise 1.59

The graph below is the graph of the function $f(x) = A \sin x + B$ for some A and B. Find these numbers.



Exercise 1.60

Half of the graph of an even function is shown below; provide the other half:



Exercise 1.61

Half of the graph of an odd function is shown above; provide the other half.

2. Exercises: Limits and continuity

Exercise 2.1

Plot the graph of the function $y = f(x)$, where x is the income (in thousands of dollars) and $f(x)$ is the tax bill (in thousands of dollars) for the income of x, which is computed as follows: no tax on the first \$10,000, then 5% for the next \$10,000, and 10% for the rest of the income. Investigate its limits and continuity.

Exercise 2.2

Explain why the limit $\lim\limits_{x\to 0} \sin\frac{1}{x}$ does not exist.

Exercise 2.3

(a) State the ε-δ definition of limit. (b) Use the definition to prove that $\lim\limits_{x\to 0} x^2 = 0$.

Exercise 2.4

(a) State the definition of limit. (b) Use the definition to prove that $\lim\limits_{x\to 0} x^3 \neq 3$.

Exercise 2.5

(a) State the definition of an infinite limit. (b) Use the definition to prove that $\lim\limits_{x\to +\infty} x^3 = +\infty$.

Exercise 2.6

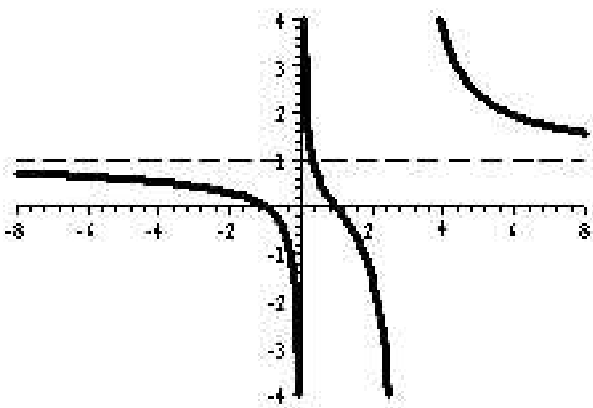
Give an example of a function with two vertical asymptotes: $x = 0$ and $x = 2$.

Exercise 2.7

Give an example of a function with a horizontal asymptote: $y = -1$, and a vertical asymptote: $x = 2$.

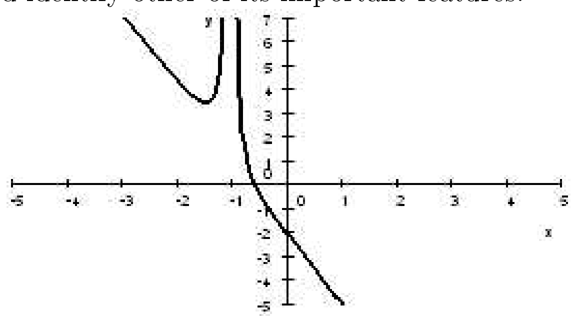
Exercise 2.8

Identify all important features of this graph:



Exercise 2.9

Express the asymptotes of this function as limits and identify other of its important features:



Exercise 2.10

True or false: "If f is continuous on (a, b), then f is bounded on (a, b)"?

Exercise 2.11

True or false: "If f is continuous on $[a, b]$, then f is bounded on $[a, b]$"?

Exercise 2.12

True or false: "If f is continuous on $[a, b)$, then f is bounded on $[a, b)$"?

Exercise 2.13

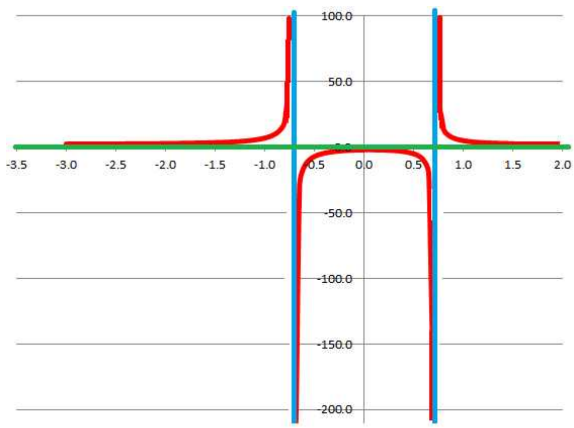
True or false: "If f is continuous on $[a, \infty)$, then f is bounded on $[a, \infty)$"?

Exercise 2.14

True or false: "Every function is bounded on a closed bounded interval"?

Exercise 2.15

The graph of f is given below. It has asymptotes. Describe them as limits. Hint: use both $+\infty$ and $-\infty$.



Exercise 2.16

A house has 4 floors and each floor has 7 windows. What was the year when the doorman's grandmother died?

Exercise 2.17

Illustrate with plots (separately) functions with the following behavior: (a) $f(x) \to +\infty$ as $x \to 1$; (b) $f(x) \to -\infty$ as $x \to 2^+$; (c) $f(x) \to 3$ as $x \to -\infty$.

Exercise 2.18

Given $f(x) = -(x-3)^4(x+1)^3$. Find the leading term and use it to describe the long term behavior of the function.

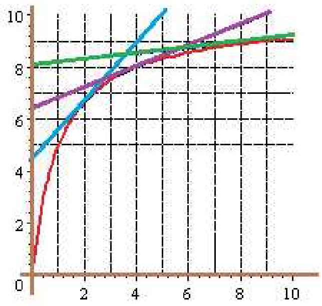
Exercise 2.19

(a) State the Intermediate Value Theorem. (b) Give an example of its application.

3. Exercises: Derivatives

Exercise 3.1

Three straight lines are shown below. What is so special about them? Find their slopes.



Exercise 3.2

(a) Suppose during the first 2 seconds of its flight an object progressed from point $(0,0)$ to $(1,0)$ to $(2,0)$. What was its average velocity and average acceleration? (b) What if the last point is $(1,1)$ instead?

Exercise 3.3

Suppose t is time and x is the price of bread. What can you say about its dynamics? Be as specific as possible.

t	x	y	z
0.00	1.00	1.00	4.00
0.20	0.83	1.36	4.03
0.40	0.71	1.64	4.07
0.60	0.63	1.84	4.09
0.80	0.56	1.96	4.07
1.00	0.50	2.00	4.00
1.20	0.45	1.96	3.87
1.40	0.42	1.84	3.67
1.60	0.38	1.64	3.41
1.80	0.36	1.36	3.07
2.00	0.33	1.00	2.67
2.20	0.31	0.56	2.19
2.40	0.29	0.04	1.63

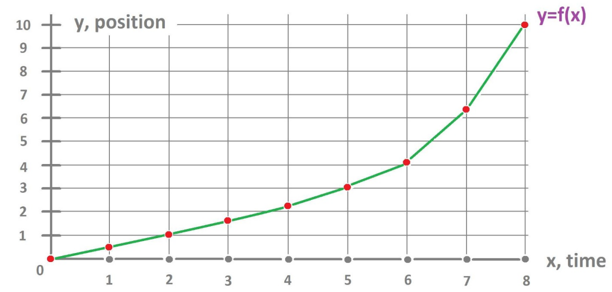
Exercise 3.4

Find the difference quotients for the function given by the following data:

x	$y=f(x)$
-1	2
1	2
3	3
5	3
7	-2
9	5

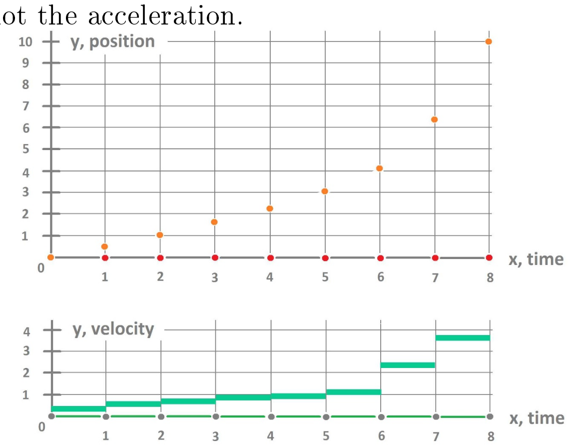
Exercise 3.5

Plot the graph of the average velocity for the following position function:



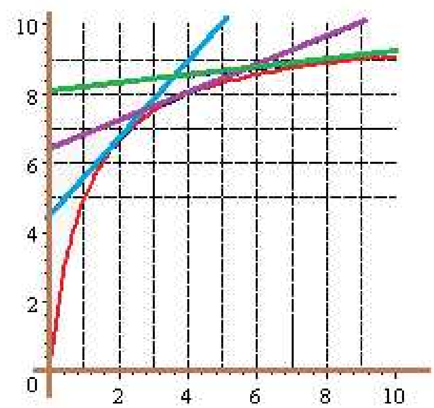
Exercise 3.6

The position and the velocity are plotted below. Plot the acceleration.



Exercise 3.7

Each of these straight lines are drown through two point of the graph. What do they tell us about the function?



Exercise 3.8

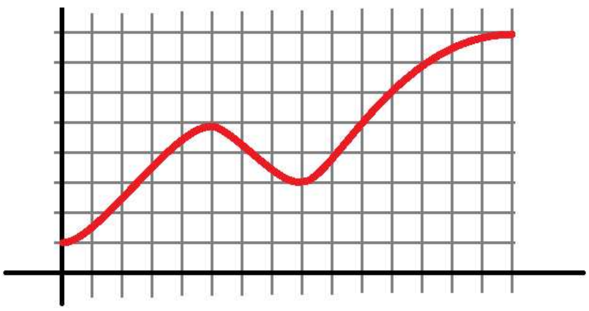
From the definition, compute the average rate of change for the function $f(x)=x^2+1$ at $a=2$ with $h=0.2$ and $h=0.1$. Explain the difference.

Exercise 3.9

(a) Compute the average rate of change for the function $f(x) = 3x^2 - x$ at $a = 1$ and $h = .5$. (b) Find the equation of the secant to the graph of $y = f(x)$ corresponding to this average rate of change.

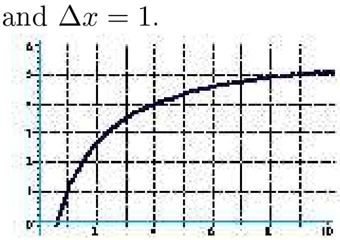
Exercise 3.10

The graph of a function $f(x)$ is given below. Estimate the values of the difference quotient $\frac{\Delta f}{\Delta x}$ for $x = 0, 4,$ and 6 and $\Delta x = 0.5$.



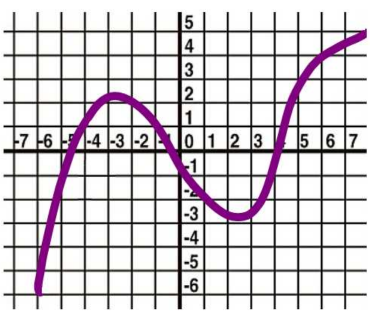
Exercise 3.11

The graph of a function $f(x)$ is given below. Estimate the values of the difference quotient for $x = 2, 4, 9$ and $\Delta x = 1$.



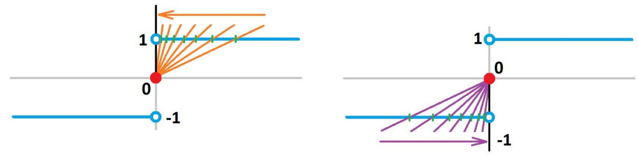
Exercise 3.12

The graph of a function f is given below. Estimate the values of the difference quotient $\dfrac{\Delta f}{\Delta x}$ for $x = 1$ and $\Delta x = 2,\ 1,\ 0.5$.



Exercise 3.13

The secant line of the sign function are shown below. What do they tell you about the differentiability of the function at $x = 0$?



Exercise 3.14

You have received the following email from your boss: "Tim, Look at the numbers in this spreadsheet. This stock seems to be inching up... Does it? If it does, how fast? Thanks. – Tom". Describe your actions.

Exercise 3.15

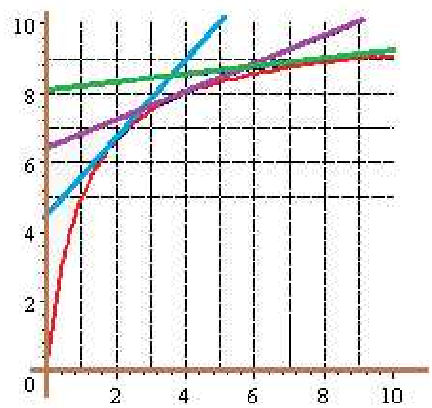
If two functions are equal, do their derivatives have to be equal too?

Exercise 3.16

Find the tangent line through the point $(2, 1)$ to the graph of the function the derivative of which is e^{x^2}.

Exercise 3.17

What do these straight lines tell us about the function?



Exercise 3.18

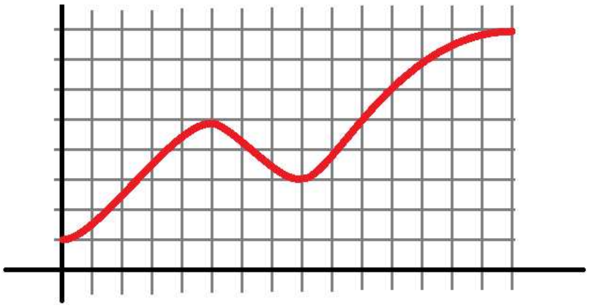
(a) State the definition of the derivative of a function at point a. (b) Provide a graphical interpretation of the definition.

Exercise 3.19

From the definition, compute the derivative of $f(x) = x^2 + 1$ at $a = 2$.

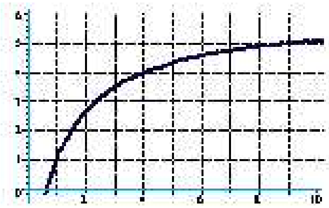
Exercise 3.20

The graph of a function $f(x)$ is given below. Estimate the values of the derivative $f'(x)$ for $x = 0, 4,$ and 6.



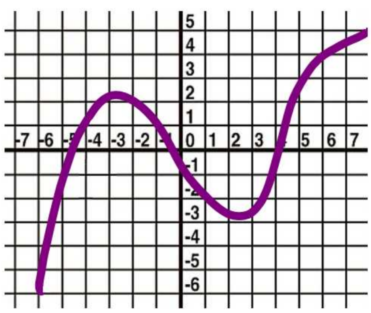
Exercise 3.21

The graph of a function $f(x)$ is given below. Estimate the values of the derivative $f'(x)$ for $x = 2, 4$, and 9.



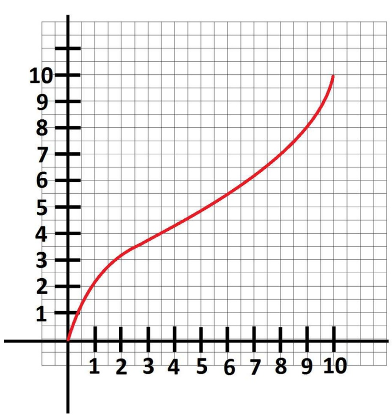
Exercise 3.22

The graph of a function f is given below. Estimate the values of the derivative f' for $x = 0$ and $x = 4$. Show your computations.



Exercise 3.23

The graph of a function $f(x)$ is given below. Estimate the values of the derivative $f'(x)$ for $x = 1, 3$, and 6.



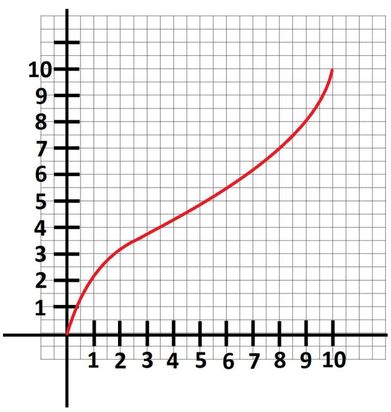
4. Exercises: Models

Exercise 4.1

The population of a city declines by 10% every year. How long will it take to drop to 50% of the current population?

Exercise 4.2

The function $y = f(x)$ shown below represents the location (in miles) of a hiker as a function of time (in hours). Sketch the hiker's velocity as the difference quotient.

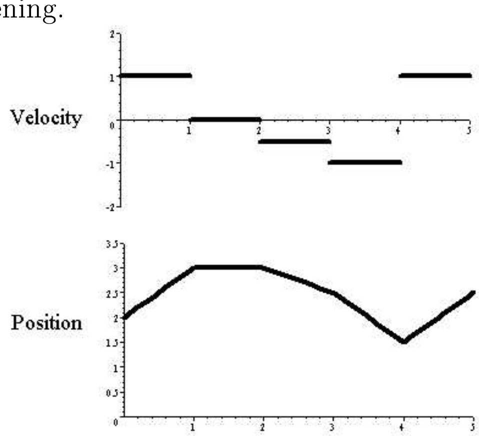


Exercise 4.3

The velocity of the object at time t is given by $v(t) = 1 + 3t^2$. If at time $t = 1$ the object is at position $x = 4$, where is it at time $t = 0$?

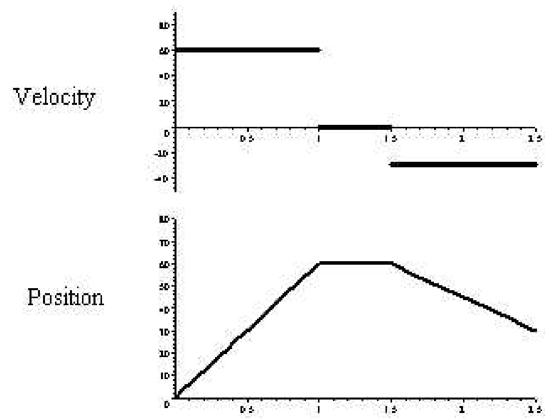
Exercise 4.4

The graphs of the velocity and the position of a moving object are shown below. Describe what is happening.



Exercise 4.5

The graphs of the velocity and the position of a moving object are shown below. Describe what is happening.



Exercise 4.6

Suppose the altitude, in meters, of an object is given by the function $t^2 + t$, where t is time, in seconds. What is the velocity when the altitude is 12 meters?

Exercise 4.7

The velocity of the object at time t is given by $v(t) = 1 + e^t$. If at time $t = 0$ the object is at $x = 2$, where is it at time $t = 1$?

Exercise 4.8

The acceleration of an object at time t is given by $a(t) = 3t$. If at time $t = 1$ the velocity of object is at $v(1) = -1$, what is it at time $t = 0$?

Exercise 4.9

Suppose $s(t)$ represents the position of a particle at time t and $v(t)$ its velocity. If $v(t) = \sin t - \cos t$ and the initial position is $s(0) = 0$, find the position $s(1)$.

Exercise 4.10

Suppose the speed of a car was growing continuously following the rule $55 + 5t$ per hour, where t is the number of hours passed since it was 250 miles away from a city. How far is it from the city after 3 hours of driving towards it?

Exercise 4.11

Let x represent the time passed since the car left the city. The table below tells for what values of x the velocity and the acceleration of the car are positive, negative, or zero. Let $f(x)$ represent the

distance of the car from the city. Sketch the graph of f.

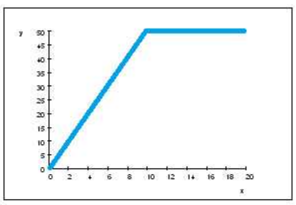
x	velocity	acceleration
0	0	$+$
1	$+$	$-$
2	0	$-$
3	$-$	$-$

Exercise 4.12

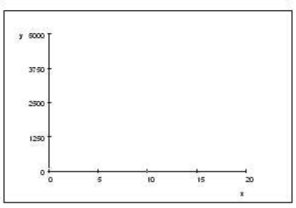
The height of the ball (in feet) t seconds after it is thrown is given by $f(t) = -16t^2 + 8t + 6$. Explain the meaning of the numbers $-16, 8, 6$.

Exercise 4.13

The graph of the velocity of a car is given below. Plot the graph of the function representing the distance of the car from the starting point.



Velocity



Distance

Exercise 4.14

Suppose the speed of a car was changing continuously following the rule $60 - t^2$ per hour, where t is the number of hours passed since noon. Find the average speed of the car between 1 pm and 3 pm.

Exercise 4.15

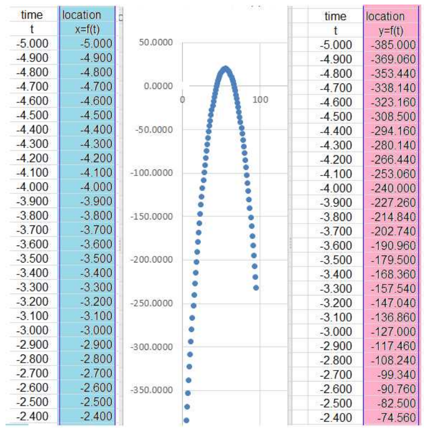
Suppose the altitude, in meters, of an object is given by the function

$$y = t^2 + t, \ t \geq 0,$$

where t is time, in sec. What is the velocity when the altitude is 12 meters?

Exercise 4.16

Find the initial conditions of a free falling object from this data:



time t	location x=f(t)
-5.000	-5.000
-4.900	-4.900
-4.800	-4.800
-4.700	-4.700
-4.600	-4.600
-4.500	-4.500
-4.400	-4.400
-4.300	-4.300
-4.200	-4.200
-4.100	-4.100
-4.000	-4.000
-3.900	-3.900
-3.800	-3.800
-3.700	-3.700
-3.600	-3.600
-3.500	-3.500
-3.400	-3.400
-3.300	-3.300
-3.200	-3.200
-3.100	-3.100
-3.000	-3.000
-2.900	-2.900
-2.800	-2.800
-2.700	-2.700
-2.600	-2.600
-2.500	-2.500
-2.400	-2.400

time t	location y=f(t)
-5.000	-385.000
-4.900	-369.060
-4.800	-353.440
-4.700	-338.140
-4.600	-323.160
-4.500	-308.500
-4.400	-294.160
-4.300	-280.140
-4.200	-266.440
-4.100	-253.060
-4.000	-240.000
-3.900	-227.260
-3.800	-214.840
-3.700	-202.740
-3.600	-190.960
-3.500	-179.500
-3.400	-168.360
-3.300	-157.540
-3.200	-147.040
-3.100	-136.860
-3.000	-127.000
-2.900	-117.460
-2.800	-108.240
-2.700	-99.340
-2.600	-90.760
-2.500	-82.500
-2.400	-74.560

5. Exercises: Information from the derivatives

Exercise 5.1

Find all local maxima and minima of the function $f(x) = x^3 - 3x - 1$.

Exercise 5.2

(a) Analyze the first and second derivatives of the function $f(x) = x^4 - 2x^2$. (b) Use part (a) to sketch its graph of f.

Exercise 5.3

Suppose the functions that follow are differentiable. (a) Finish the statement "If $h'(x) = 0$ for all x in (a, b), then...". (b) Finish the statement "If $f'(x) = g'(x)$ for all x in (a, b), then...".

Exercise 5.4

Sketch the graph of the function $f(x) = \sqrt{x}e^{-x}$. Justify the graph by studying the derivatives of f.

Exercise 5.5

(1) State Rolle's Theorem and illustrate it with a sketch. (b) Quote and state the theorem(s) necessary to prove it. (c) What theorem follows from it?

Exercise 5.6

Sketch the graph of the function given below. Provide justification for each feature of the graph:

$$f(x) = \frac{x^2 + 7x + 3}{x}.$$

Exercise 5.7

(a) State the Mean Value Theorem. (b) Verify that the function $f(x) = \frac{x}{x+2}$ satisfies the hypotheses of the theorem on the interval $[1, 4]$.

Exercise 5.8

Sketch the graph of the function $f(x) = x^4 - x^2$. Provide justification for each feature of the graph.

Exercise 5.9

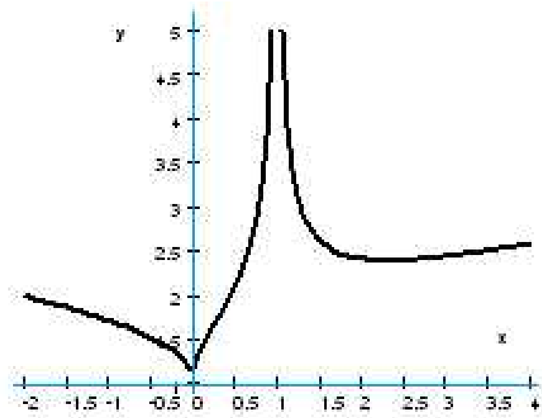
Find global maxima and minima of the function, $f(x) = x^3 - 3x$ on the interval $[-2, 10]$.

Exercise 5.10

Find the local maximum and minimum points of the function $f(x) = x^3 - 3x$.

Exercise 5.11

The graph of function f is given below. (a) At what points is f continuous? (b) At what points does the derivative of f exist?



Exercise 5.12

Indicate which the following statements below is true or false (no proof necessary):

1. If the function f is increasing, then so is f^{-1}.
2. The exponential function has an asymptote.
3. If $f'(c) = 0$, then c is a local maximum or a local minimum of f.
4. If a function is differentiable then it is continuous.
5. If two functions are equal, their derivatives are also equal.
6. If two functions are equal, their antiderivatives are also equal.

6. Exercises: Computing derivatives

Exercise 6.1

Suppose $f(1) = 3$ and $f'(1) = 2$. Use this information to fill in the blanks:

$$\left(f^{-1}(\quad)\right)' =$$

Exercise 6.2

Differentiate this:

$$g(t) = t\cos t \sin t\,.$$

Exercise 6.3

Differentiate:
$$\frac{\ln(\sin x)}{x}\,.$$

Exercise 6.4

Compute the derivative of $f(x) = e^{x^2+3x}$.

Exercise 6.5

Evaluate $\frac{d}{dx}\left(\sin x \cdot e^{x+1}\right)$.

Exercise 6.6

Evaluate $\frac{d}{dx}\left(\cos t + e^{t}\right)$. Hint: watch the variables.

Exercise 6.7

Evaluate $\frac{dy}{dx}$ for $y = \sin e^{2x}$.

Exercise 6.8

Evaluate the derivative of of $f(x) = xe^{\sin x}$.

Exercise 6.9

Suppose $f'(1) = 2$, $g'(2) = 3$, and $h'(1) = 6$, where $h = g \circ f$. What is $f(1)$?

Exercise 6.10

Is it possible that both $F(x)$ and $F(2x)$ are both antiderivatives of some function f?

Exercise 6.11

Is $\sin x + 3x$ an antiderivative of $\cos x^2$?

Exercise 6.12

Is it possible that both $F(x)$ and $F(2x)$ are both antiderivatives of some non-zero function f?

Exercise 6.13

Evaluate the derivative of $f(x) = x^2e^x$.

Exercise 6.14

Find the second derivative of $h(x) = x^2 + x + 1$. What does it tell you about the shape of the graph of f?

Exercise 6.15

Find the second derivative of $h(x) = 2x^{\pi}$.

Exercise 6.16

Compute the derivative of $f(x) = \ln(3x + 2)$.

Exercise 6.17

Find the second derivative of $h(x) = xe^x$.

Exercise 6.18

Find the derivatives of the functions: (a) $3x^e + e^{\pi}$, (b) $7\ln x + (1/x) - \ln 2$.

Exercise 6.19

Evaluate $\frac{dy}{dx}$ for
$$y = \sqrt{e^x}\,.$$

Exercise 6.20

Find the slopes of the tangent lines to the ellipse $x^2 + 2y^2 = 1$ at the points where it crosses the diagonal line $y = x$.

Exercise 6.21

Evaluate $\frac{dy}{dx}$ for $y = \sin\cos(-x)$.

Exercise 6.22

Suppose $x\sin y + y^2 = x$. Find $\frac{dy}{dx}$.

7. Exercises: Riemann integral

Exercise 7.1

(a) State the Fundamental Theorem of Calculus. (b) Use part (a) to evaluate

$$\int_{-1}^{1} \sin \frac{x}{3}\, dx\,.$$

Exercise 7.2

(a) State the Fundamental Theorem of Calculus. (b) Use part (a) to evaluate

$$\frac{d}{dx}\int_{0}^{x} e^{t^2}\, dt\,.$$

Exercise 7.3

(a) Make a sketch of the left-end Riemann sums for $\int_0^1 \sqrt{x}\, dx$ with $n = 4$ intervals. (b) State the algebraic properties of the Riemann integral.

Exercise 7.4

Given $f(x) = x^2 + 1$, write (but do not evaluate) the Riemann sum for the integral of f from -1 to 2 with $n = 6$ and left ends as sample points. Make a sketch.

Exercise 7.5

Provide the definition of the definite integral via its Riemann sums. Make a sketch.

Exercise 7.6

The Fundamental Theorem of Calculus includes the formula $\int_a^b f(x)\, dx = F(b) - F(a)$. (a) State the whole theorem. (b) Provide definitions of the items appearing in the formula.

Exercise 7.7

(a) State the definition of the definite integral $\int_a^b f(x)\, dx$ and illustrate the construction with a sketch. (b) Use the definition to justify that

$$\int_a^b cf(x)\, dx = c\int_a^b f(x)\, dx$$

for a constant c.

Exercise 7.8

Suppose

$$\int_0^1 f\, dx = 2,\ \int_0^4 f\, dx = 0,\ \int_1^2 f\, dx = 2\,.$$

Find

$$\int_1^3 f\, dx,\ \int_0^1 (f(x)+3)\, dx,\ \int_2^4 f\, dx\,.$$

Exercise 7.9

Suppose a function is defined by:

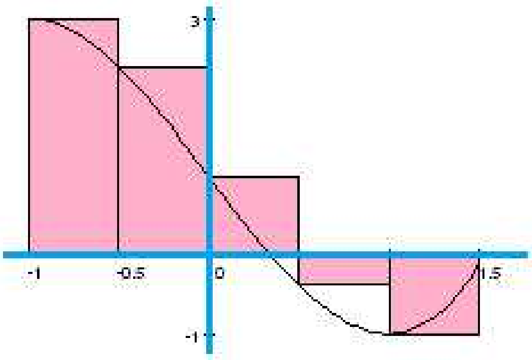
$$F(x) = \int_2^x f\, dx.$$

Find, in terms of F, the following:

$$\int_0^4 f\, dx,\ \int_1^2 f\, dx,\ \int_0^{-1} f\, dx,\ \int_1^2 (f(x)-1)\, dx\,.$$

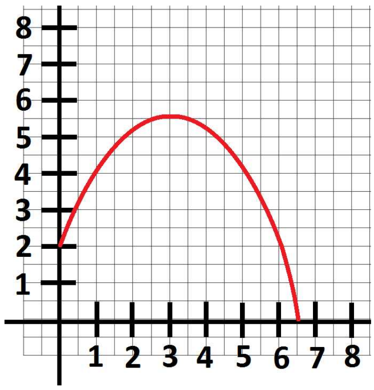
Exercise 7.10

Evaluate the Riemann sum of f below on the interval $[-1, 1.5]$ with $n = 5$. What are its sample points? What does it estimate?



Exercise 7.11

Write (don't evaluate) the left-end Riemann sum of the integral $\int_0^5 f(x)\, dx$ for function f shown below with $n = 5$ intervals.



Exercise 7.12

Write the formula and illustrate with a sketch the left-end Riemann sum L_4 of the integral $\displaystyle\int_1^3 f(x)\,dx$ for the function plotted above.

Exercise 7.13

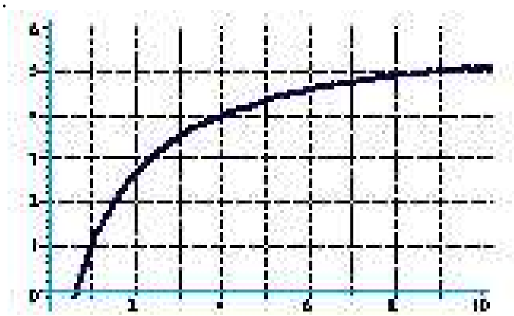
Write the mid-point Riemann sum that approximates the integral $\displaystyle\int_0^1 \sin x\,dx$ within .01.

Exercise 7.14

Set up the Riemann sum for the area of the circle of radius R as the area between two curves, provide an illustration and the integral formula. Evaluate for extra 5 points.

Exercise 7.15

Let $I = \displaystyle\int_2^8 f\,dx$. (a) Use the graph of $y = f(x)$ below to estimate L_4, M_4, R_4. (b) Compare them to I.



Exercise 7.16

Complete the following statements:

- $(f(x)\cdot x^2)' = f'(x)\cdot x^2 + ...$
- $\displaystyle\int x^{-1}\,dx = ...$
- $\displaystyle\int f'(x)\,dx = ...$
- $\displaystyle\int u\,dv = uv...$
- $u = \cos t \implies du = ...$

Exercise 7.17

Suppose that F is an antiderivative of a differentiable function f. If F is increasing on $[a,b]$, what can you say about f?

8. Exercises: Integration

Exercise 8.1

Execute the following substitution in the integral (don't evaluate the resulting integral):

$$\int \sqrt{\cos x + \sin x}\, dx, \quad u = \sin x\,.$$

Exercise 8.2

Suppose $s(t)$ represents the position of a particle at time t and $v(t)$ its velocity. If $v(t) = \sin t - \cos t$ and the initial position is $s(0) = 0$, find the position $s(1)$.

Exercise 8.3

Evaluate

$$\int e^{3x}\, dx\,.$$

Exercise 8.4

Evaluate

$$\int_1^2 (e^x + \sqrt{x} + x^{-1})\, dx\,.$$

Exercise 8.5

Evaluate

$$\int e^{x^2} 2x\, dx\,.$$

Exercise 8.6

Evaluate

$$\int 2x \sin 5x\, dx\,.$$

Exercise 8.7

Evaluate

$$\int_1^3 e^{t+1}\, dx\,.$$

Hint: watch the variables.

Exercise 8.8

Calculate:

$$\int \left(e^{\sin x^2 + 77}\right)'\, dx\,.$$

Exercise 8.9

Evaluate the integral by substitution

$$\int x e^{x^2}\, dx\,.$$

Exercise 8.10

Find all antiderivatives of the following function: $f(x) = e^{-x}$.

Exercise 8.11

Find the antiderivative F of the function $f(x) = 3x^2 - 1$ satisfying the initial condition $F(1) = 0$.

Exercise 8.12

Evaluate the integral

$$\int_0^1 x^3\, dx\,.$$

Exercise 8.13

Evaluate:

$$\int x^2\, dx - \int x^2\, dx\,.$$

Exercise 8.14

Evaluate:

$$\int x^{-2}\, dx - \int x^{-2}\, dx\,.$$

Exercise 8.15

Integrate by parts:

$$\int 3x e^{-x}\, dx\,.$$

Exercise 8.16

Use the table of integrals to evaluate:

$$\int \sin^{-1} 2x\, dx\,.$$

Exercise 8.17

Evaluate:

$$\int_0^1 \frac{1}{2x}\, dx\,.$$

Exercise 8.18

Use substitution to evaluate the integral:

$$\int_0^\pi \sin x \ \cos^2 x \, dx \,.$$

Exercise 8.19

Integrate by parts:

$$\int x(\ln x)^2 \, dx \,.$$

Exercise 8.20

Use substitution to evaluate the integral:

$$\int_0^1 \frac{1}{\sqrt{4-x^2}} \, dx \,.$$

Exercise 8.21

Use the table of integrals to evaluate:

$$\int x^2(\sqrt{x^2-4} - \sqrt{x^2+9}) \, dx \,.$$

Exercise 8.22

Evaluate

$$\int x \sin x \, dx \,.$$

Exercise 8.23

Use substitution $u = 1+x^2$ to evaluate the integral

$$\int \sqrt{1+x^2} x^5 \, dx \,.$$

Exercise 8.24

Use substitution to evaluate the integral:

$$\int_0^\pi \sin x \ \cos^2 x \, dx \,.$$

Exercise 8.25

Evaluate the improper integral:

$$\int_1^\infty \frac{1}{2x} \, dx \,.$$

Exercise 8.26

Find the antiderivative F of the function $f(x) = e^x + x$ satisfying the initial condition $F(0) = 1$.

9. Exercises: Applications of integrals

Exercise 9.1

The region bounded by the graphs of $y = \sqrt{x}$, $y = 0$, and $x = 1$ is revolved about the x-axis. Find the surface area of the solid generated.

Exercise 9.2

A chord of a circle is a straight line segment whose end-points lie on the circle. Find the average length of a chord perpendicular to the diameter. What about parallel?

Exercise 9.3

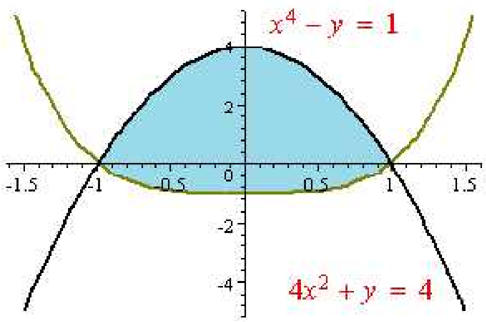
Find the average length of a segment in a square parallel to (a) the base, (b) the diagonal.

Exercise 9.4

Find (by integration) the length of a circle of radius r.

Exercise 9.5

Find the area enclosed by the curves below:



Exercise 9.6

Find the area of the region bounded by $y = x^2 - 1$ and $y = 3$.

Exercise 9.7

Suppose f is an integrable function. (a) Show that f is also odd then $\displaystyle\int_{-a}^{a} f\,dx = 0$. (b) Suggest a related formula for an even f.

Exercise 9.8

Find the centroid of the region bounded by the curves $y = x^2,\ y = 1$.

Exercise 9.9

Find the x-coordinate of the center of mass of the region between $y = x^2$ and $y = x^3$.

Exercise 9.10

Find the volume of a right circular cone of radius R and height h by any method you like.

Exercise 9.11

Compute the average area of the cross section of the sphere of radius 1.

Exercise 9.12

Find the center of mass of the region below $y = 2x$ for $0 \le x \le 1$.

Exercise 9.13

The volume of a solid is the integral of the areas of its cross-sections. Explain and justify using Riemann sums.

Exercise 9.14

The region bounded by the graphs of $y = x^2 + 1,\ y = 0,\ x = 0$ and $x = 1$ is revolved about the x-axis. Find the volume area of the solid generated.

Exercise 9.15

The region bounded by the graphs of $y = x^2 + 1,\ y = 0,\ x = 0,$ and $x = 1$ is revolved about the y-axis. Find the volume area of the solid generated.

Exercise 9.16

An aquarium 2 m long, 1 m wide, and 1 m deep is full of water. Find the work needed to pump half of the water out of the aquarium (the density of water is 1000 kg/m^3).

Exercise 9.17

Find the area of the surface of revolution around the x-axis obtained from $y = \sqrt{x},\ 4 \le x \le 9$.

Exercise 9.18

Find the centroid of the region bounded by the curves $y = 4 - x^2,\ y = x + 2$.

Exercise 9.19

Find the area of the region bounded by $y = x^2 - 1$ and $y = 3$.

Exercise 9.20

Find the area under the graph of the function $f(x) = e^x$ from $x = -1$ to $x = 1$.

Exercise 9.21

Find the average value of the function $f(x) = 2x^2 - 3$ on the interval $[1, 3]$.

Exercise 9.22

Find the area of the region bounded by $y = \sqrt{x}$, the x-axis, and the lines $x = 1$ and $x = 4$.

10. Exercises: Parametric curves and related

Exercise 10.1

Describe the motion of a particle with position (x, y), where

$$x = 2 + t\cos t,\ y = 1 + t\sin t,$$

as t varies within $[0, \infty)$.

Exercise 10.2

Suppose the parametric curve is given by

$$x = \cos 3t,\ y = 2\sin t.$$

Set up, but do not evaluate, the integrals that represent (a) the arc-length of the curve, (b) the area of the surface obtained by rotating the curve about the x-axis.

Exercise 10.3

Suppose curve C is the graph of function $y = f(x)$. (a) Find a parametric representation of C. (b) Find a parametric representation of C that goes from right to left.

Exercise 10.4

Find all points on the curve

$$x = \cos 3t,\ y = 2\sin t$$

where the tangent is either horizontal or vertical.

Exercise 10.5

Sketch the following parametric curve:

$$x = |\cos t|,\ y = |\sin t|,\ -\infty < t < +\infty.$$

Describe the curve and the motion.

Exercise 10.6

Sketch the following parametric curves:

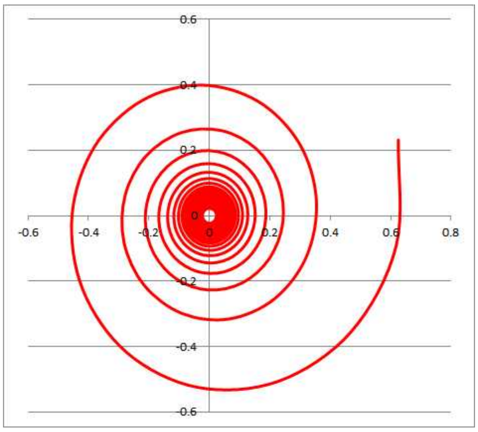
- $x(t) = \dfrac{1}{t},\ y(t) = \sin t,\ t > 0$
- $x = \cos t,\ y = 2$
- $x = 1/t,\ y = 1/t^2,\ t > 0$

Exercise 10.7

(1) Sketch the parametric curve $x = \cos t,\ y = \sin 2t$. (2) The curve intersects itself. Find the angle of this intersection.

Exercise 10.8

Find an equation of the spiral converging to the origin as below:

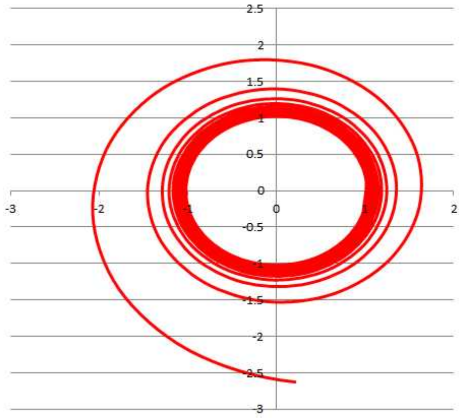


Exercise 10.9

Plot this entire parametric curve: $x = \sin t,\ y = \cos 2t$.

Exercise 10.10

Find a parametric representation of a curve similar to the one below, a spiral wrapping around a circle. What about one that is wrapping from the inside? (no proof necessary):



Exercise 10.11

Given a parametric curve $x = \sin t,\ y = t^2$. Find the line(s) tangent to the curve at the origin.

Exercise 10.12

Find a parametric representation of a curve that looks like the figure eight or a flower (no proof necessary).

Exercise 10.13

Sketch the parametric curve:

$$x = t^2 - 1,\ y = 2t^2 + 3\,.$$

Exercise 10.14

Suppose that the following parametric curve represents the motion of an object on the plane:

$$x = 3t - 1,\ y = t^2 - 1\,.$$

(a) When does the object cross the x-axis? (b) When does the object cross the y-axis?

Exercise 10.15

Represent as a parametric curve the rotation of a rod of length 2 that makes one full turn every 3 seconds.

Exercise 10.16

One circle is centered at $(0, 0)$ and has radius 1. The second is centered at $(3, 3)$. What is the radius of the second if the two circles touch?

Exercise 10.17

Represent in polar coordinates these points given by their Cartesian coordinates: (a) $(1, 2)$; (b) $(-1, -1)$; (c) $(0, 0)$.

Exercise 10.18

Represent in Cartesian coordinates these points given by their polar coordinates: (a) $\theta = 0,\ r = -1$; (b) $\theta = \pi/4,\ r = 2$; (c) $\theta = 1,\ r = 0$.

Exercise 10.19

(a) Represent the following complex number in the standard form: $(2+3i)(-1+2i)$. Indicate the real and imaginary parts. (b) Find its module and argument.

Exercise 10.20

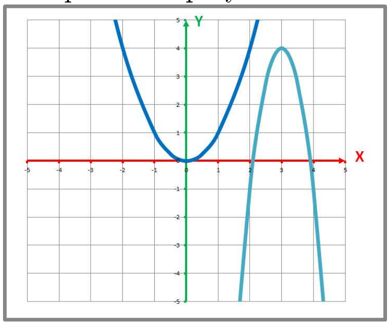
Simplify $(1+i)^2$.

Exercise 10.21

(a) Find the roots of the polynomial $x^2 + 2x + 2$. (b) Find its x-intercepts. (c) Find its factors.

Exercise 10.22

What can you say about the imaginary parts of the roots of these quadratic polynomials?



Exercise 10.23

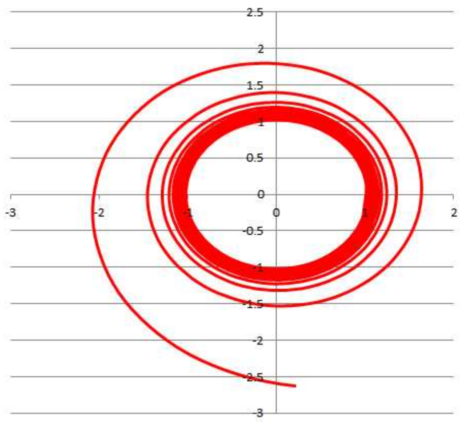
Plot the curve $r = 2\cos(3\theta)$. For 5 extra points, find the line(s) through the origin tangent to the curve.

Exercise 10.24

(a) Plot these points in polar coordinates: $(r, \theta) = (0, 1),\ (1, 0),\ (1, \pi),\ (2, 3\pi)$. (b) Sketch these three polar curves: $r = 1,\ \theta = 0,\ r = \theta$.

Exercise 10.25

Find a polar representation of a curve similar to the one below, a spiral wrapping around a circle. What about one that is wrapping from the inside? (no proof necessary):



Exercise 10.26

Indicate if the following statements are true or false:

1. In polar coordinates, $(1, \pi/2)$ and $(-1, -\pi/2)$ represent the same point.
2. The curve $r = 3 + \cos\theta$ passes through the origin.
3. The curve $r = \cos 2\theta$ is closed.
4. The curve $r = 1 + \cos\theta$ is bounded.
5. The graph of $r = \theta^2$ can be represented as a parametric curve in Cartesian coordinates.

6. In polar coordinates, $A = (1, \pi/2)$ and $B = (-1, -\pi/2)$ represent the same point.

7. The slope of the polar curve $r = 0$ is equal to 0.

8. The graph of the curve $r = \cos 2\theta$ is a spiral.

9. The parametric curve $x = t^2,\ y = \sin t$ is bounded.

10. The graph of $r = \theta^2$ can be represented as a parametric curve in Cartesian coordinates.

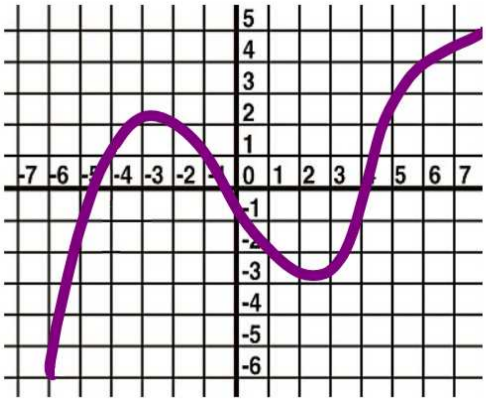
11. Exercises: Several variables

Exercise 11.1

Draw a few level curves of the function $f(x,y) = x^2 + y$.

Exercise 11.2

The graph of function $y = g(x)$ of *one* variable is shown below. Suppose now that $z = f(x,y) = g(x)$ is a function of *two* variables, which depends only on x, given by the same formula. Find all points where the gradient of f is equal to 0.

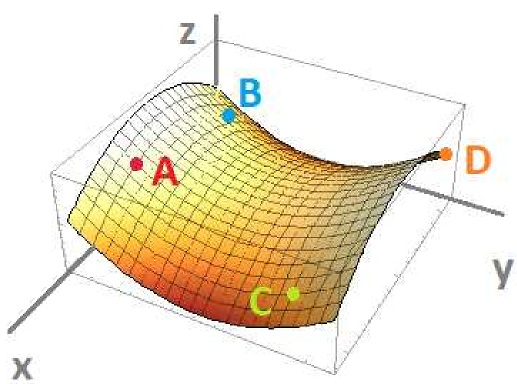


Exercise 11.3

Find all critical points of the function $f(x,y) = 2x^3 - 6x + y^2 - 2y + 7$.

Exercise 11.4

Sketch the contour (level) curves of the function shown below, along with points A, B, C, D, on the xy-plane:



Exercise 11.5

Sketch the level curves of the function $f(x,y) = 2xy + 1$ for the following values of $z = -1, 0, 1, 2$.

Exercise 11.6

Show that the limit doesn't exist:

$$\lim_{(x,y)\to(0,0)} \frac{xy}{x^2+y^2}\,.$$

Exercise 11.7

Draw the contour map (level curves) of the function $f(x,y) = e^{y/x}$. Explain what the level curves are.

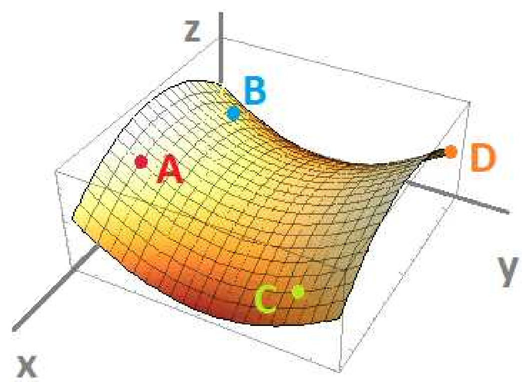
Exercise 11.8

Sketch the graph of a function of two variables $z = f(x,y)$ the derivatives of which have the following signs:

$$f_x > 0,\ f_{xx} > 0,\ f_y < 0,\ f_{yy} < 0\,.$$

Exercise 11.9

The graph of a function of two variables $z = f(x,y)$ is given below along with four points on the graph. Sketch the gradient for each on a separate xy-plane:

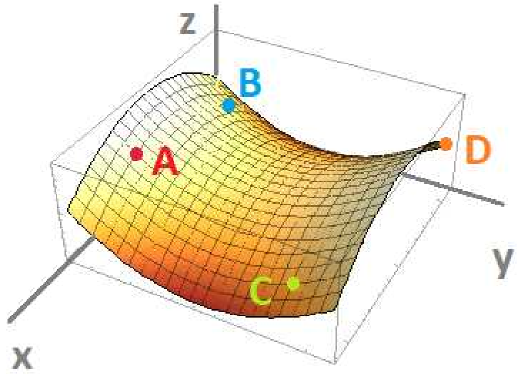


Exercise 11.10

Find the gradient of the function $f(x,y) = x^2y^{-3}$ at the point $(1,1)$. Use this information to sketch the graph of f in the vicinity of this point. Explain.

Exercise 11.11

The graph of a function of two variables $z = f(x,y)$ is given below along with four points on the graph. Provide the signs (positive or negative) of the partial derivatives of f at these points. For example, $\dfrac{\partial f}{\partial x} < 0$ at point A.

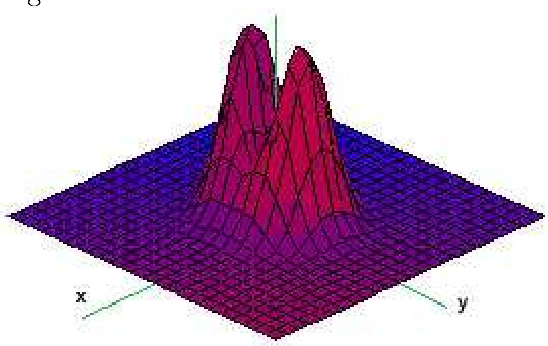


Exercise 11.12

Find all critical points of the function $f(x,y) = 2x^3 - 6x + y^2 - 2y + 7$.

Exercise 11.13

Make a sketch of contour (level) curves for the following function:



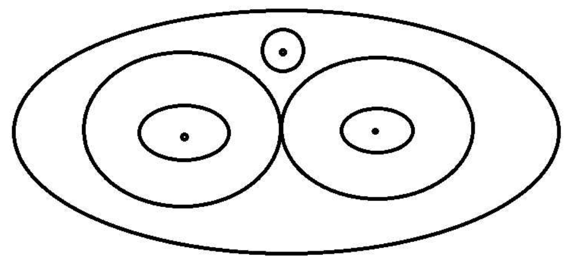
Exercise 11.14

The wave heights h in the open sea depend on the speed v of the wind and the length of time t that the wind has been blowing at that speed. Values of the function $h = f(v,t)$ are recorded in the table below. Estimate the rate of change of h with respect to v when $v = 40$ and $t = 15$. Show your computations.

$v \backslash t$	15	20	25
30	16	17	18
40	25	28	31
50	36	40	45

Exercise 11.15

The contour (level) curves for a function are given below. They are equally spaced. Sketch a possible graph that produced it and describe it.



Exercise 11.16

Draw the contour map (level curves) of the following function of two variables:

- $g(x,y) = \ln(x+y)$
- $f(u,v) = uv$
- $h(x,y) = 2x - 3y + 7$
- $z = x^2 + y^2$

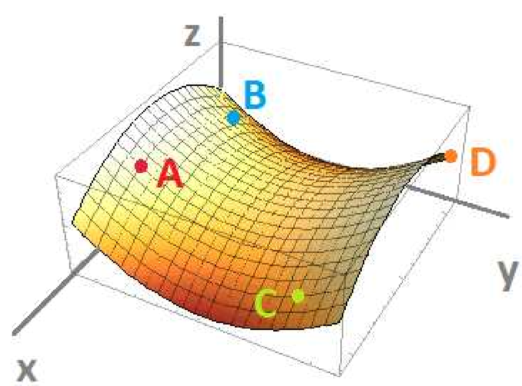
Exercise 11.17

Sketch the graph of a function of two variables $z = f(x,y)$ the derivatives of which have the following signs:

	1	2	3	4
f_x	+	+	−	+
f_{xx}	−	+	+	−
f_y	−	−	+	+
f_{yy}	−	+	−	+

Exercise 11.18

The graph of a function of two variables $z = f(x,y)$ is given below along with a point on the graph: 1. A, 2. B, 3. C, 4. D. Determine the signs of the derivatives f_x, f_{xx}, f_y, f_{yy} at that point:



12. Exercises: Sequences and series

Exercise 12.1

Present the first 5 terms of the sequence:

$$a_1 = 1, \quad a_{n+1} = -(a_n + 1)\,.$$

Exercise 12.2

Represent in sigma notation:

$$-1 - 2 - 3 - 4 - 5 - \ldots - 10\,.$$

Exercise 12.3

Find the following sum:

$$-1 - 2 - 3 - 4 - 5 - \ldots - 10\,.$$

Exercise 12.4

Find the sequence of sums of the following sequence:

$$-1, 2, -4, 8, -5, \ldots$$

Exercise 12.5

Compute $\displaystyle\sum_{n=1}^{4} n^2$.

Exercise 12.6

Show that $\frac{n}{n+1}$ is an increasing sequence. What kind of sequence is $\frac{n+1}{n}$? Give examples of increasing and decreasing sequences.

Exercise 12.7

Find the next item in each list:

1. $7, 14, 28, 56, 112, \ldots$
2. $15, 27, 39, 51, 63, \ldots$
3. $197, 181, 165, 149, 133, \ldots$

Exercise 12.8

A pile of logs has 50 logs in the bottom layer, 49 logs in the next layer, 48 logs in the next layer, and so on, until the top layer has 1 log. How many logs are in the pile?

Exercise 12.9

In the beginning of each year, a person puts \$5000 in a bank that pays 3% compounded annually. How much does he have after 15 years?

Exercise 12.10

An object falling from rest in a vacuum falls approximately 16 feet the first second, 48 feet the second second, 80 feet the third second, 112 feet the fourth second, and so on. How far will it fall in 11 seconds?

Exercise 12.11

Evaluate the limit if it exists:

$$\lim_{n\to\infty} \frac{(-1)^n}{n}\,.$$

Exercise 12.12

Evaluate the limit: $\displaystyle\lim_{n\to\infty} \ln\left(\frac{n}{n+1}\right)$.

Exercise 12.13

Give an example of a sequence for each of the following: (a) $a_n \to 0$ as $n \to \infty$, (b) $a_n \to 1$ as $n \to \infty$, (c) $a_n \to +\infty$ as $n \to \infty$, (d) a_n diverges but not to infinity.

Exercise 12.14

Write a formula for the nth term of the sequence:

$$-\frac{1}{2}, \frac{3}{4}, -\frac{7}{8}, \frac{15}{16}, -\frac{31}{32}, \ldots$$

Exercise 12.15

Explain why the limit $\displaystyle\lim_{n\to\infty} \sin n$ does not exist.

Exercise 12.16

(a) State the Squeeze Theorem. (b) Give an example of its application.

Exercise 12.17

Present the first 5 terms of the sequence and tell if it is convergent:

$$a_1 = 1, \quad a_{n+1} = (a_n - 1)^2\,.$$

Exercise 12.18

(a) State the definition of the sum of a series. (b) Use (a) to prove the Sum Rule.

Exercise 12.19

Find the sum of the series

$$\sum_{n=0}^{\infty} \frac{(-1)^n + 2}{3^n}.$$

Exercise 12.20

Test the following series for convergence (including absolute/conditional):

$$\sum \frac{(-1)^{n-1}}{(1.1)^n}.$$

Exercise 12.21

Apply the Integral Test to show that the p-series with $p = 1/3$ diverges.

Exercise 12.22

Test the following series for convergence (including absolute/conditional):

$$\sum (-1)^{2n} \frac{1}{n^n}.$$

Exercise 12.23

Test the following series for convergence (including absolute/conditional):

$$\sum \frac{n^{1/2}}{n^2 - 1}.$$

13. Exercises: Power series and approximations

Exercise 13.1

Estimate the coefficients of the Taylor polynomial T_1 of order 1 centered at $a = 1$ of the function f shown above. Provide a formula for this T_1. What about T_2; what is the sign of the coefficient of the next term that appears in T_2?

Exercise 13.2

What degree Taylor polynomial one would need to approximate $e^{.01}$ within .001? (Answers may vary and yours doesn't have to be perfect but it has to be justified.)

Exercise 13.3

(a) State the definition of absolute convergence. (b) Give an example of a series that converges but not absolutely.

Exercise 13.4

What degree Taylor polynomial one would need to approximate $\sin(-0.01)$ within 0.001? Explain the formula:

$$E_n \leq K_{n+1}\frac{|x-a|^{n+1}}{(n+1)!}$$

and why you can choose $K_{n+1} = 1$.

Exercise 13.5

Find the interval of convergence of the series:

$$\sum \frac{(x-2)^n}{n}\,.$$

Exercise 13.6

Explain how functions are represented by power series and how they both are differentiated. Demonstrate on $f(x) = e^x$.

Exercise 13.7

Find the Taylor polynomial of degree 4 that would help to approximate $e^{1.01}$.

Exercise 13.8

Find the radius and the interval of convergence of the series

$$\sum \frac{2(x+1)^n}{n^2}\,.$$

Exercise 13.9

Find the Taylor series centered at $a = 1$ of the function $f(x) = x^4$.

Exercise 13.10

Find the radius and the interval of convergence of the series:

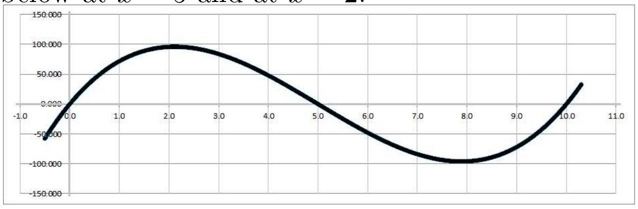
$$\sum \frac{(x-1)^n}{\sqrt{n}2^n}\,.$$

Exercise 13.11

Find the Taylor polynomial $T_2(x)$ of order 2 centered at $a = \pi$ of the function $f(x) = \sin^2 x$.

Exercise 13.12

Give a linear approximation of the function plotted below at $x = 5$ and at $x = 2$:



Exercise 13.13

Use linear approximation of $f(x) = \sin x$ to estimate $\sin .02$.

Exercise 13.14

Find the linear approximation of $f(x) = \ln x$ at $a = 1$. Use it to estimate $\ln .99$.

Exercise 13.15

Find the linear approximation of $f(x) = \sqrt{x}$ at $a = 1$. Use it to estimate $\sqrt{1.1}$.

Exercise 13.16

Find the linear approximation of $f(x) = \sin 3x$ at $a = 0$. Use it to estimate $\sin -.02$.

Exercise 13.17

Find the linear approximation of $f(x) = \sqrt{1+3x}$ at $a = 0$. Use it to estimate $\sqrt{1.03}$.

Exercise 13.18

Find the linear approximation to estimate $\sqrt[3]{26.9}$.

Exercise 13.19

Find the quadratic approximation of $f(x) = x^{1/3}$ at $a = 1$. Use it to estimate $1.1^{1/3}$.

Exercise 13.20

Use linear approximation to estimate $\sin \pi/2$.

Exercise 13.21

Use quadratic approximation to estimate $\sin \pi/4$.

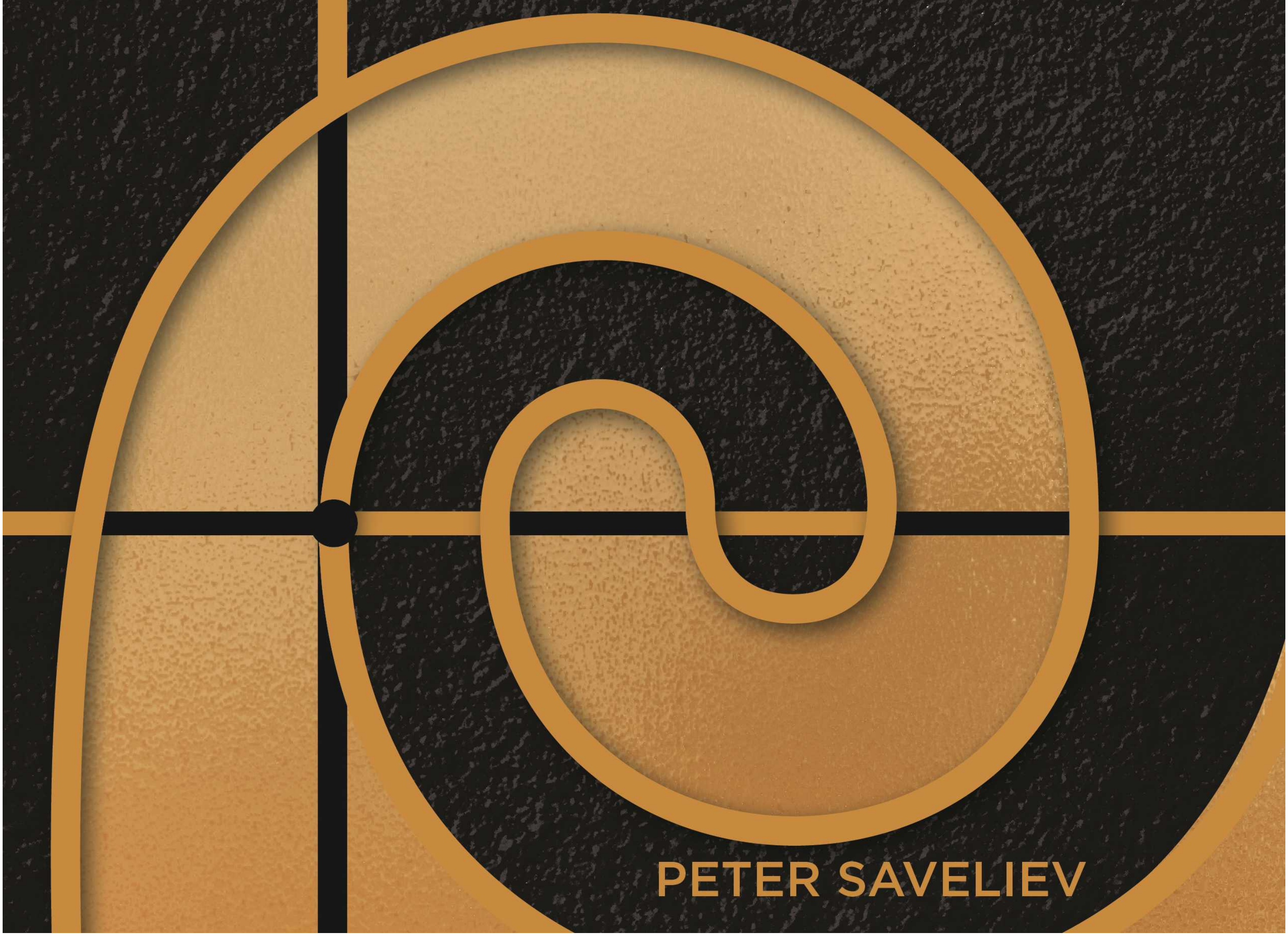
Exercise 13.22

Give an example of a function the best linear approximation of which coincide with the constant approximation.

Next...

CALCULUS ILLUSTRATED

VOLUME 4: CALCULUS IN HIGHER DIMENSIONS



PETER SAVELIEV

Index